Databases and Information Systems
BalticDB&IS‘2012

Lithuanian Academy of Sciences
Vilnius University
Lithuanian Computer Society

Databases and Information Systems

Tenth International Baltic Conference on Databases and Information Systems

Local Proceedings, Materials of Doctoral Consortium

Vilnius, Lithuania
July 8-11, 2012

Edited by

Albertas Čaplinskas
Vilnius University
Lithuania

Gintautas Dzemyda
Vilnius University
Lithuania

Audronė Lupeikienė
Vilnius University
Lithuania

Olegas Vasilecas
Vilnius Gediminas Technical University
Lithuania

Vilnius 2012

UDK 004.65(474.5)(06)
Da372

Databases and Information Systems. Tenth International Baltic Conference on Databases and Information Systems. Local Proceedings, Materials of Doctoral Consortium. A. Čaplinskas, G. Dzemyda, A. Lupeikienė, O. Vasilecas (Eds.). Vilnius: Žara, 2012, 264 p.

This book consists of three independent parts. The first part contains the short communications and accepted full research papers presented at the 10th Jubilee International Baltic Conference on Databases and Information Systems (BalticDB&IS'2012) but not included in the post-proceedings published by IOS Press Frontiers in Artificial Intelligence and Applications series. The second part contains the abstracts of the papers published in post-proceedings. Finally, the third part contains the materials of Doctoral Consortium. Short communications and full research papers discuss the wide spectrum of research topics including databases, data mining and optimisation in information systems, business modelling, cloud computing and e-service, software evaluation and testing, decision support systems, norms and reasoning, information system engineering tools and techniques, and advanced e-learning environments and technologies. In the materials of Doctoral Consortium doctoral students present the current state of their research.

Cover design: A. Gedgaudas

Typesetting: manuscripts were formatted by conference technical staff

ISBN 978-9986-34-274-8

ORGANISED BY

Lithuanian Academy of Sciences

Vilnius University,
Institute of Mathematics and Informatics

Lithuanian Computer Society

SPONSORS

Research Council of Lithuania

Algoritmų Sistemos, Ltd.

Information Technologies Institute

Conference Committee

Steering Committee

Juris BORZOVS, University of Latvia, Latvia
Janis BUBENKO, Stockholm University, Sweden
Albertas ČAPLINSKAS, Vilnius University, Lithuania
Jānis Grundspeņķis, Riga Technical University, Latvia
Hele-Mai HAAV, University of Technology, Estonia
Ahto KALJA, Tallinn University of Technology, Estonia
Mārīte KIRIKOVA, Riga Technical University, Latvia
Audronė LUPEIKIENĖ, Vilnius University, Lithuania
Arne SØLVBERG, Norwegian University of Science and Technology, Norway
Olegas VASILECAS, Vilnius Gediminas Technical University, Lithuania

General Chair

Gintautas DZEMYDA, Vilnius University, Lithuania

Program Committee Co-Chairs

Albertas ČAPLINSKAS, Vilnius University, Lithuania
Olegas VASILECAS, Vilnius Gediminas Technical University, Lithuania

Publishing Co-Chair

Audronė LUPEIKIENĖ, Vilnius University, Lithuania

Programme Committee

Irina ASTROVA, Tallinn University of Technology, Estonia
Marko BAJEC, University of Ljubljana, Slovenia
Romas BARONAS, Vilnius University, Lithuania
Guntis BĀRZDIŅŠ, University of Latvia, Latvia
András BENCZÚR, Eötvös Loránd University, Hungary
Jānis BIČEVSKIS, University of Latvia, Latvia
Mária BIELIKOVÁ, Slovak University of Technology in Bratislava, Slovakia
Juris BORZOVS, University of Latvia, Latvia
Boštjan BRUMEN, University of Maribor, Slovenia
Dumitru Dan BURDESCU, University of Craiova, Romania
Rimantas BUTLERIS, Kaunas University of Technology, Lithuania
Vytautas ČYRAS, Vilnius University, Lithuania
Paweł CZARNUl, Gdansk University of Technology, Poland
Valentina DAGIENĖ, Vilnius University, Lithuania
Marlon DUMAS, University of Tartu, Estonia
Dalė DZEMYDIENĖ, Mykolas Romeris University, Lithuania

Additional Reviewers

Natalia GARANINA, Russia
Nada BAJNAID, Saudi Arabia
Arne KOSCHEL, Germany
Kristina SMILGYTĖ, Lithuania
Lovro ŠUBELJ, Slovenia
Milos RADOVANOVIC, Serbia
Tomaž HOVELJA, Slovenia
Simon VRHOVEC, Slovenia
Tarmo ROBAL, Estonia
Riina MAIGRE, Estonia
Ateeq KHAN, Germany
Norbert SIEGMUND, Germany
Azeem LODHI, Germany
Iwona DUBIELEWICZ, Poland
Lech TUZINKIEWICZ, Poland
Bogumiła HNATKOWSKA, Poland

Doctoral Consortium Advising Committee

Coordinator

Prof. Dalė DZEMYDIENĖ, Mykolas Romeris University, Vilnius University Institute of Mathematics and Informatics, Lithuania

Advisors:

Prof. Jānis GRUNDSPEŅĶIS, Riga Technical University, Latvia
Prof. Hele-Mai HAAV, University of Technology, Estonia
Prof. Mārīte KIRIKOVA, Riga Technical University, Latvia
Prof. Henrikas PRANEVIČIUS, Kaunas University of Technology, Lithuania

Coordinators of Special Sections

Olga KURASOVA, Vilnius University (Section: Data Mining and Optimisation in IS)
Valentina DAGIENĖ, Vilnius University (Section: Advanced e-Learning Environments and Technologies)

Local Organising Committee Chair

Saulius MASKELIŪNAS, Lithuanian Computer Society

Local Organising Committee

Bronius JASKELEVIČIUS, Litnuanian Academy of Sciences
Viktor MEDVEDEV, Vilnius University
Jolanta MILIAUSKAITĖ, Vilnius University
Laima PALIULIONIENĖ, Vilnius University
Sandra SVANIDZAITĖ, Vilnius University
Marius ŠAUČIŪNAS, Vilnius University
Dalia ŠUKVIETIENĖ, Lithuanian Computer Society
Aidas ŽANDARIS, Lithuanian Computer Society

Preface

The series of Baltic DB&IS biennial conferences was started in 1994. The first conference was held in Trakai, Lithuania. It was initiated by Prof. Janis A. Bubenko Jr. (Royal Institute of Technology, Stockholm, Sweden) and Prof. Arne Sølvberg (Norwegian University of Science and Technology, Norway) and organised by Institute of Mathematics and Informatics (Lithuania), Vilnius University (Lithuania), Vilnius Technical University (Lithuania). This time was a hard time for Baltic research community. After the crash of Soviet Union the cooperation with the research centres in Moscow, Leningrad, Kiev, and Novosibirsk was broken. The horizontal cooperation between research centres in Lithuania, Latvia and Estonia was weak developed. We had almost no touch with the research centres in West countries and even in East European countries. The idea of Janis A. Bubenko jr. and Arne Sølvberg to organise Baltic Workshops on databases and information systems was indeed a great idea. It helped to consolidate research communities in Baltic countries, to establish contacts with research centres in West and East European countries, and even to renew contacts with the researchers in Russia. These workshops have become a real incentive for development of regional research and boost of international cooperation.

Janis A. Bubenko jr. and Arne Sølvberg helped also to receive founding and even trained us in West European conference organization technologies. Despite the fact that Tallinn, Riga and Vilnius had big experience in organizing large soviet conferences, this experience was almost not applicable in organizing conferences according to West standards. Besides, the time was very hard. It is enough to mention that during the conference in Trakai the conference venue and diner room were protected by armed safeguard.

In year 2000, the international workshop was transformed into international conference. Two years earlier, the selected papers of the conference started to be published by international publishers. The detail statistic data are presented in the table below.

The analysis of this table shows not only positive but also some negative trends. Starting from year 2008, the geography of participants contracted significantly. It seems that it happened not only for reasons of world-wide financial crisis, but also for changes in state scientific policies. In our opinion, the trend to ignore international conferences and to accent publications only in high-quality international scientific journals is not perspective and even malign.

Another negative trend is the permanently enlarged gap between research communities and industry. If first conferences had significant industrial tracks and intensive round table discussions, in three recent conferences the participation of industry partners is less than minimal. In our opinion, this is the evidence that, at least in Baltic countries, the state policy oriented to encourage the cooperation among research centres and industry is not effective. Despite these negative trends, we remain optimistic and believe that Baltic DB&IS biennial conferences reborn for its new life.

In the final, we express our warmest thanks to all authors who contributed to the 10th Jubilee Conference, to the members of international Programme Committee, additional reviewers, and to the organizing team. We also express our very special thanks and deep gratitude to all our sponsors. Last, but not the least, we also thank all the participants of the conference.

Year/ Venue/ Status	**Proceedings**	**PC**		**Submissions**		
		Number of members	**Number of countries**	**Number of submissions**	**Number of countries**	**Accepted**
1994 Trakai Workshop	Local	16	5		11	43 (Australia-1, Belgium - 1, Estonia-+ , Finland-2, France-3, Latvia-5, Lithuania-7, Norway-3, Sweden-10, UK-3, USA-2)
1996 Tallinn Workshop	Local	25	14	55	19	36 (Australia-3.5, Estonia-5, Finland-3.5, France-5, Germany-2.75, Israel-0.25, Latvia-7, Lithuania-1, Malaysia-1, Netherlands – 1, Russia-1, Sweden-3, Switzerland-1, UK-3, USA-1)
1998 Riga Workshop	Local	31	16	46	17	39 (Australia-0.5, Brasil-1, Denmark-1, Estonia-1, Finlandia-4.5, France-1, Germany-2, Hungary-1, Korea-1, Latvia-10, Lithuania-5.5, Malaysia-1, Netherlands-1, Poland-3, Sweden-1, UK-1.5, USA-1)
2000 Vilnius Workshop	Local + **Kluwer Academic Publishers** (selected papers)	37	18	60		39 (Belarus-1, Brasil-2, Estonia-5, Finland-4, France-1, Germany-9.33, Japan-1, Latvia-5.8, Lithuania-14, Norway-1, Poland-5, Russia-3, Spain-1, Sweden-0.33, UK-2.03, Ukraine-1, USA-2.5, Yugoslavia-1)
2002/ Tallinn **Conference**	Local + Kluwer Academic Publishers (selected papers)	33	18	60	23	41 (Australia-0.5, Czech Republic-1, China-1, Denmark-2, Estonia-11, Germany-7, Italy-1, Japan-1, Latvia-11, Lithuania-3, New Zealand-1, Norway-1.5, Poland-3, Portugal-1, Russia-3, Spain-1, Sweden-1, UK-1, USA-1)
2004 Riga Conference	Local + IOS Press (selected papers)	35	14	59	17	35 (Czech Republic-3, Estonia-3, Finland-1, France-1, Germany-5.5, Latvia-4, Lithuania-4, New Zealand-1,

Year/ Venue/ Status	**Proceedings**	**PC**		**Submissions**		
		Number of members	**Number of countries**	**Number of submissions**	**Number of countries**	**Accepted**
						Norway-1, Poland-3, Russia-3.5, Spain-2, Sweden-1, UK-1, USA-1)
2006 Vilnius Conference	Local + IOS Press (selected papers) + IEEE (selected papers)	87	35	84	21	48 (Australia-1, Austria-2, Belgium-0.3, Brazil-0.5, China-1, Denmark-0.5, Estonia-2, France-5, Germany-2, Italy-1.5, Latvia-17, Lithuania-11.8, Mexico-0.5, Netherlands-0.7, New Zealand-2, Portugal-1, Russia-1, Spain-0.5, UK-1.7, USA-1)
2008 Tallinn Conference	Local + IOS Press (selected papers)	43	22	43	12	29 (Austria-1, Estonia-8, Finland-1, Germany-2, Latvia-13, Lithuania-5, Poland-1, Russia-1, UK-1)
2010 Riga Conference	Local + IOS Press (selected papers)	51	15			36 (Austria-0.5, Estonia-6.5, Germany-2.5, Latvia-17, Lithuania-2, Poland-1, Russia-0.5, Sweden-1)
2012 Vilnius Conference	Local + IOS Press (selected papers)	64	23	69	14	43 (Algeria-1, Canada-1, Estonia-6, France-3, Germany-3, Latvia-17, Lithuania-9.5, Russia-3, Slovenia-2, Switzerland-0.5)

July 2012

Albertas Čaplinskas
Olegas Vasilecas

Contents

RESEARCH PAPERS AND SHORT COMMUNICATIONS

Databases

Data Mining and Optimisation in IS

Business Modelling

Cloud Compuring and E-Service

Software Evaluation and Testing

Decision Support Systems, Norms, and Reasoning

IS Engineering Tools and Techniques

Advanced e-learning Environments and Technologies

ABSTRACTS

MATERIALS OF DOCTORAL CONSORTIUM

Part 1. Research Papers and Short Communications

Workload-based Heuristics for Evaluation of Physical Database Architectures

Andreas LÜBCKE [a], Martin SCHÄLER [a], Veit KÖPPEN [a] and Gunter SAAKE [a]
[a] *School of Computer Science,*
Otto-von-Guericke-University Magdeburg, Germany,
{andreas.luebcke,martin.schaeler,veit.koeppen,gunter.saake}@ovgu.de

Abstract. Database systems are widely used in different application domains. Therefore, it is difficult to decide which database management system meets the requirements of a certain application at most. This observation is also true for scientific and statistical data management, due to new application and research fields. New requirements are often implied to data management while discovering unknown research and applications areas. That is, heuristics and tools do not exist to select an optimal database management system. In previous work, we proposed a decision framework based on application workload analyses. Our framework supports application performance analyses by mapping and merging workload information to patterns. In this paper, we present heuristics for performance estimation to select an optimal database management system for a given application. We show that these heuristics improve our decision framework by complexity reduction without loss of accuracy.

Keywords. Heuristics, storage architecture, design, performance, query processing

Introduction

Database systems (DBS) are pervasively for almost each branch of business activity. Therefore, DBS have to manage different requirements for heterogeneous application domains. New data management approaches are developed (e.g., NoSQL-DBMSs [9,14], MapReduce [12,13], Cloud Computing [3,16,7], etc.) to make the growing amount of data[1] manageable for special application domains. We argue, these approaches are developed for special applications and need a high degree of expert knowledge for usage, administration, and optimization. However, we focus our observations to relational database management systems (DBMSs) in this paper. Relational DBMSs are commonly used DBS for highly diverse applications and besides relational DBMS are well-know to many IT-affine people.

Relational DBMSs[2] are developed to manage data of daily business and reduce paper trails of companies (e.g., finance institutions) [2]. This approach dominates more and more the way of data management that we today know as online transaction processing (OLTP). Nowadays, faster and more accurate forecasts for revenues and expenses are not enough anymore. A new application domain evolves that focuses on analysis of data to support business decisions. Codd et al. [8] defines this type of data analysis as

[1] Consider the data explosion problem [22,26].

[2] In the following, we use the term DBMS synonymously for relational DBMS.

online analytical processing (OLAP). Consequently, two disjunctive application domains for relational data management exist with different scopes, impacts, and limitations (cf. Section 1).

In recent years, business application have a high demand for solutions that support tasks from both OLTP and OLAP [15,21,28,31,33,34], thus coarse heuristics for typical OLTP and/or OLAP applications have become obsolete (e.g., data warehouses without updates always perform best on column-oriented DBMSs). Nevertheless, new approaches, which we mention above, also show impacts and limitations (e.g., in-Memory-DBMS only, focus on real-time or dimension updates, etc.), such that we argue there is no DBMS that fits for OLTP and OLAP in all application domains. Heuristics and current approaches for physical design and query optimization only consider a certain architecture[3] (e.g., design advisor [36] and self-tuning [10] for row-oriented DBMSs or equivalent for column-oriented DBMSs [19,30]). That is, the decision for a certain architecture has to be done beforehand. Consequently, there is no approach that neither advices physical design spanning different architectures for OLTP, OLAP, and mixed OLTP/OLAP workloads nor that estimates which architecture is optimal to process a query/database operation.

However, we refine obsolete heuristics for physical design of DBS (e.g., heuristics from classical OLTP domain). We consider OLTP, OLAP, and mixed application domains for physical design. We present heuristics that propose the usage of row-oriented DBMSs (row stores) or column-oriented DBMSs (column stores) under certain circumstances. Furthermore, we present heuristics for query execution or rather for processing (relational) database operations on column and row stores. Our heuristics show which query type and/or database operation performs better on a particular architecture[4] and how single database operations affect performance of a query or a workload. We derive our heuristics from experiences in workload analyses with the help of our decision model, presented in [23].

In the following sections, we consider the main differences between row- and column store as well as the advantages and disadvantages for column stores. Section 2 addresses our heuristics for physical design of DBMSs concerning different storage architectures (i.e., row or column store). In Section 3, we present heuristics for query processing on row and column stores. Section 4 gives an overview of related research. Finally, we summarize our discussions and give an outlook in Section 5.

1. Column or Row Store: Assets and Drawbacks

In previous work, we already discussed differences between column and row stores according to different subjects [25,24,23]. We can summarize our major observations in Table 1. Of course, column stores architecture also has disadvantages. We just name them because they are mostly contradictory to row store architecture advantages. Column stores perform worse on update operations and concurrent non-read-only data access due to partitioned data, thus on frequent updates and consistency checks tuple reconstructions cause notable cost.

Finally, row stores can outperform column stores in their traditional application domain nor vice versa. Other researchers confirm our consideration that one architecture

[3] We use the term architecture synonymously for storage architecture.

[4] The term architecture refers to row- and column-oriented database architecture.

cannot sustain the other architecture in their native domain (e.g., by simulating architecture through partitioned schema [4]).

Table 1. Advantages of Column Stores

Requirement	CS Property	Comment
Disc space	Reduced	Aggressive compression (e.g., optimal compression per data type)
Data transfer	Less	More data fits in main memory
Data processing	Compressed and decompressed	Does not work for each compression nor for all operations
OLAP I/O	No	Neither for aggregations nor column operations
Parallelization	For inter- and intra-query	Not for ACID-transaction with write operations
Vector operations	Fast	Easily adaptable

2. Heuristics on Physical Design

Physical design of DBSs is important as long as DBMSs exist. We do not only restrict our considerations to a certain architecture. Further, we consider a set of heuristics that can be used to forecast which architecture is more suitable for a given application.

Some existing rules still have their validity. First, pure OLTP applications perform best on row stores. Second, classic OLAP application with an ETL (extract, transform, and load) process or (very) rare updates are satisfied with column stores[5]. In the following, we consider a more exciting question. In which situation one architecture outperforms the other one and in which case they perform nearly equivalent.

OLTP For OLTP workloads, we just recommend to use row stores as we do it for decades. A column store does not achieve competitive performance except column-store architecture will significantly change.

OLAP In this domain one might suspect a similar situation as for OLTP workloads. However, this is not true in general. We are aware that column stores outperform row stores for many applications and/or queries in this domain; that is, for aggregates and access as well as processing of a few columns. In most cases, column stores are most suitable for applications in this domain. Nevertheless, there exist complex OLAP queries where column stores lose their advantages (cf. Section 1). For these complex queries, row stores can achieve competitive results, even if they consume more memory. These queries have to be considered for architecture selection because they critically influence the physical design estimation even more if there is a significant amount of these queries in workload.

OLTP/OLAP In these scenarios, your physical design strongly depends on the ratio between updates, point queries, and analytical queries. Our experience is that column stores perform about 100-times slower on OLTP-transactions (updates, inserts, etc.) than row stores. This fact is even worse because we do not even consider concurrency (e.g., ACID); that is, we make this observation in single-user execution on transactions. Assuming transaction and analytical queries in average take the same time, we state that one transaction only occurs every 100 queries (OLAP). The fact that analytical queries

[5] Note: Not all column stores support updates just ETL.

last longer than a single iteration leads us to a smaller ratio. Our experience shows that 10 executions of analytical queries on a column store are of greater advantage than the loss by a single transaction. If you have a smaller ration than 10:1 (analyses/Tx) then we cannot give a clear statement. We recommend using a row store in this situation or you know beforehand that the ratio will change to more analytical queries. If the ratio falls under this ratio for a column store and is not a temporary change then a system change is appropriate. So, in mixed workloads it is all about the ratio of analytical queries to transactions. Note that the ratio 100:1 and 10:1 can change considering OLAP-query type. We address this issue in Section 3.

Our heuristics can be used as a guideline for architecture decisions for certain applications. That is, we select the most suitable architecture for an application and afterwards use existing approaches (like IBM's advisor [36]) to tune physical design of a certain architecture. If workload and DBSs are available for analysis, we emphasize to use our decision model [23] to calculate an optimal architecture for a defined workload. The presented heuristics for physical design extent our decision model to reduce calculation costs (i.e., solution room is pruned). Additionally, heuristics make our decision model available for scenarios where only restricted information is available.

3. Heuristics on Query Execution

In the following, we present heuristics for query execution on complex or hybrid DBS. We assume a DBS that supports column- and row-store functionality or a complex environment with at least two DBMSs containing data redundant whereby at least one DBMS is setup for each OLTP and OLAP. In the following, we only discuss query processing heuristics for OLAP and OLTP/OLAP workloads due to our assumptions above and the fact that there is no competitive alternative to row stores for OLTP.

OLAP In this domain, we face a huge amount of data that generally is not frequently updated. Column stores are able to significantly reduce the amount of data due to aggressive compression. That is, more data can be loaded into main memory and I/O between storage and main memory is reduced. We state that this I/O reduction is the major benefit of column stores. We made the experience that row stores perform worse on many OLAP queries because row stores drop performance due to fact that CPUs often are idle while waiting for I/O from storage. Moreover, row stores read data that is not required due to the physical design of data. In our example from TPC-H benchmark [32], one can see that only a few columns of the lineitem relation have to be accessed (cf. Listing 1). In contrast to column stores, row stores have to access the complete relation to answer this query. We state, most OLAP query fit into this pattern. We recommend using column store functionality to answer this type of OLAP queries as long as they only access a minority of columns from relation for aggregation and predicate selection.

```
from    lineitem
where   l_shipdate >= date '1994-01-01'
    and l_shipdate < date '1994-01-01' + interval '1' year
    and l_discount between .06 - 0.01 and .06 + 0.01
    and l_quantity < 24;
```

Listing 1. TPC-H query Q6

Complex Queries OLAP is often used for complex analyses, thus queries become more complex too. These queries describe complex issues and/or produce large business reports. Our example (Listing 2) from the TPC-H benchmark could be part of a report (or a more complex query). Complex queries access and/or aggregate many tuples, that is, nearly the complete relation has to be read. This implies a number of tuple reconstructions that significantly reduce the performance of column stores. Hence, row stores can achieve competitive performance because nearly the complete relations have to be accessed. Our example shows another reason for a number of tuple reconstructions: group operations. Tuples have to be reconstructed before aggregating groups. Other reasons for significant performance reduction by tuple reconstructions can be a large number of predicate selections on different columns as well as complex joins. We argue, this complex OLAP query type can be executed on both architectures. This fact can be used to load balance queries in complex environments and hybrid DBMSs.

```
    select c_custkey, count(o_orderkey) from
        customer left outer join orders on c_custkey = o_custkey
        and o_comment not like '%special%request%'
        group by c_custkey) as c_orders (c_custkey, c_count)
group by c_count
order by custdist desc, c_count desc;
```

Listing 2. TPC-H query Q13

Mixed Workloads In mixed workload environments, our first recommendation is to split the workload into two parts OLTP and OLAP. Both parts can be allocated to the corresponding DBS with row- or column-store functionality. As we mentioned above, this split methodology can also be used for load balancing if one DBS is too busy. With this approach, we achieve competitive performance for both OLAP and OLTP because, according to our assumptions, complex systems have to have at least one DBMS of each architecture. Further, we state that a future hybrid system has to satisfy both architecture, too. Processing mixed workloads with our split behavior has two additional advantages. First, we can reduce issues with complex OLAP queries by correct allocation or dividing over both DBS. Second, we can consider time-bound parameters for queries.

As mentioned above, two integration methods for the query-processing heuristics are available. First, we propose the integration into a hybrid DBMSto decide where to optimally execute a query. That is, heuristics enable rule-based query optimization for hybrid DBMS as we know from row-store optimizer [17]. Second, we propose a global manager on top of complex environments to optimally distribute queries (as known from distributed DBMSs [27]). Our decision model analyzes queries and decides where to distribute queries to. Afterwards, queries are locally optimized by DBMS itself. Note, time-bound requirements change over time and system design determines how up-to-date data is in OLAP system (e.g., real-time load [31] is available or not). We state that such time-bound requirements can be passed as parameters to our decision model. That is, even OLAP queries can be distributed to OLTP part of the complex environment if data in OLAP part is insufficiently updated or vice versa we have to ensure analysis on most up-to-date data. Additionally, such parameters can be used to alternatively allocate high-priority queries to another DBS part if the query has to wait otherwise.

4. Related Work

Several approaches are developed to analyze and classify workloads (e.g., [18,29]). These approaches recommend tuning and design of DBS, try to merge similar tasks, etc. to improve performance of DBSs. To the present, workload-analysis approaches are limited to classification of queries to execution pattern or design estimations for a certain architecture; that is, the solution space is beforehand pruned analysis and performance estimations are done. This results in information loss due to an inappropriate reduction. With our decision model and heuristics, we propose an approach that is independent from architectural issues.

Due to success in the analytical domain, researchers devote more attention on column stores whereby focused on analysis performance [1,19] or to overcome update-problems with separate storages [15,1]. However, a hybrid system in an architectural manner does not exist. In-memory-DBMSs are developed in recent years (e.g., HyPer [21]) that satisfy requirements for mixed OLTP/OLAP workloads. Nevertheless, we state that even today not all DBSs can run in-memory (due to monetary or environmental constraints), thus we propose a more general approach.

Recommending indexes [5,10], materialized views [10], configurations [20], or physical design in general [6,37,36] are in focus of researchers. However, all approaches are limited to certain DBMSs or at least one architecture to the best of our knowledge. Our approach is situated on top of these design and tuning approaches. That is, we support a first coarse granular design and tuning for a global view on hybrid systems and utilize existing approaches to optimize locally.

In the literature, researchers compare the performance of column and row stores considering certain scenarios. Cornell and Yu [11] focus on disc-access minimization for transaction to improve query execution time. Abadi et al. [4] compare different approaches for column-oriented storage concerning analytical queries. We state, current applications demand for a combined observations of OLTP and OLAP scenarios.

In line with Zukowski et al. [35], we observe benefits to convert from NSM to DSM and vice versa during query processing. In contrast to Zukowski et al., we do not only focus on CPU trade-offs caused by tuple reconstructions. Our approach is used for both scenarios and considers overall benefit of global and local tuning.

5. Conclusion

In recent years, many new approaches as well as new requirements encourage performance of DBMSs in many application domains. Nevertheless, new approaches and requirements also increase complexity of DBMS selection, DBS design, and tuning for a certain application. We focus our considerations on OLTP, OLAP, and OLTP/OLAP workloads in general. That is, we considered which DBMS is most suitable.

Therefore, we present heuristics on design estimations and query execution for both relational storage architectures row and column stores. We use our decision model [23] to observe the performance of several applications. Consequently, we derive heuristics from our experiences while evaluating our decision model. We present these heuristics for DBS design to a priori select the most suitable DBMS and to tune this system afterwards. Our approach avoids misleading tuning if the architecture selection is wrong. Furthermore, we present heuristics on query execution for both storage architectures. On the one hand, we want to emphasize our heuristics for physical design. On the other hand, we propose these heuristics for integration in hybrid DBS whether it is a real hybrid DBMS or it is a complex system that consists of different DBMSs (at least one row and one column store). Summarizing, there is no alternative in OLTP environments to row stores, for OLAP applications we recommend to use column stores taking into account that a number of complex OLAP queries can change this recommendation, and finally in mixed OLTP/OLAP environments the focus is on the ratio of OLAP queries to OLTP transactions (again taking very complex OLAP queries into account). Our heuristics are a first step to rule-based query optimization in hybrid systems/architectures.

In future work, we will evaluate our heuristics considering standard benchmark to achieve meaningful results. After evaluation, we will implement the heuristics in our decision model, thus we achieve a design advisor for both relational storage architectures. Finally, we plan to implement a hybrid DBMSs using our heuristics for ruled-based query optimization.

Acknowledgements

This paper has been funded in part by the German Federal Ministry of Education and Science (BMBF) through the Research Program under Contract FKZ: 13N10817.

References

[1] Daniel J. Abadi. Query execution in column-oriented database systems. PhD thesis, Cambridge, MA, USA, 2008. Adviser: Madden, Samuel.

[2] Morton M. Astrahan, Mike W. Blasgen, Donald D. Chamberlin, Kapali P. Eswaran, Jim Gray, Patricia P. Griffiths, W. Frank King III, Raymond A. Lorie, Paul R. McJones, James W. Mehl, Gianfranco R. Putzolu, Irving L. Traiger, Bradford W. Wade, and Vera Watson. System R: Relational Approach to Database Management. *ACM Trans. Database Syst.* **1**(2) (1976), 97–137.

[3] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski, Gunho Lee, David A. Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. Above the Clouds: A Berkeley View of Cloud Computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, Feb 2009.

[4] Daniel J. Abadi, Samuel R. Madden, and Nabil Hachem. Column-stores vs. row-stores: How different are they really? In *SIGMOD'08* (2008), 967–980.

[5] Nicolas Bruno and Surajit Chaudhuri. To tune or not to tune? A lightweight physical design alerter. In *VLDB'06*, VLDB Endowment (2006), 499–510.

[6] Nicolas Bruno and Surajit Chaudhuri. An online approach to physical design tuning. In *ICDE'07*, IEEE (2007), 826–835.

[7] Rajkumar Buyya, Chee Shin Yeo, and Srikumar Venugopal. Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities. In *HPCC* (2008), 5–13.

[8] Edgar F. Codd, Sally B. Codd, and Clynch T. Salley. Providing OLAP to User-Analysts: An IT Mandate. *Ann ArborMichigan* (1993), 24.

[9] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber. Bigtable: A Distributed Storage System for Structured Data. In *OSDI* (2006), 205–218.

[10] Surajit Chaudhuri and Vivek Narasayya. Self-tuning database systems: A decade of progress. In *VLDB'07* (2007), 3–14.

[11] Douglas W. Cornell and Philip S. Yu. An effective approach to vertical partitioning for physical design of relational databases. *Trans. Softw. Eng.* **16**(2) (1990), 248–258.

[12] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *OSDI* (2004), 137–150.

[13] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Commun. ACM* **51**(1) (2008), 107–113.

[14] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon's highly available key-value store. In *SOSP* (2007), 205–220.

[15] Clark D. French. Teaching an OLTP database kernel advanced datawarehousing techniques. In *ICDE'97* (1997), 194–198.

[16] Ian T. Foster, Yong Zhao, Ioan Raicu, and Shiyong Lu. Cloud Computing and Grid Computing 360-Degree Compared. *CoRR*, abs/0901.0131, 2009.

[17] Goetz Graefe and David J. DeWitt. The EXODUS Optimizer Generator. In *SIGMOD'87* (1987), 160–172.

[18] Marc Holze, Claas Gaidies, and Norbert Ritter. Consistent on-line classification of DBS workload events. In *CIKM'09* (2009), 1641–1644.

[19] Stratos Idreos. Database Cracking: Torwards Auto-tuning Database Kernels. PhD thesis, 2010.

[20] Eva Kwan, Sam Lightstone, K. Bernhard Schiefer, Adam J. Storm, and Leanne Wu. Automatic database configuration for DB2 Universal Database: Compressing years of performance expertise into seconds of execution. In *BTW'03*, GI (2003), 620–629.

[21] Alfons Kemper and Thomas Neumann. HyPer: A hybrid OLTP&OLAP main memory database system based on virtual memory snapshots. In *ICDE'11* (2011), 195–206.

[22] Henry F. Korth and Abraham Silberschatz. Database Research Faces the Information Explosion. *Commun. ACM* **40**(2) (1997), 139–142.

[23] Andreas Lübcke, Veit Köppen, and Gunter Saake. A Decision Model to Select the Optimal Storage Architecture for Relational Databases. In *Proceedings of the Fifth IEEE International Conference on Research Challenges in Information Science, RCIS* (2011), 74–84.

[24] Andreas Lübcke and Gunter Saake. A Framework for Optimal Selection of a Storage Architecture in RDBMS. In *DB&IS* (2010), 65–76.

[25] Andreas Lübcke. Challenges in Workload Analyses for Column and Row Storess. In *Grundlagen von Datenbanken* (2010).

[26] Ina Naydenova and Kalinka Kaloyanova. Sparsity Handling and Data Explosion in OLAP Systems. In *MCIS* (2010), 62–70.

[27] M. Tamer Özsu and Patrick Valdurie. *Principles of Distributed Database Systems*. Springer, 3rd edition, 2011.

[28] Hasso Plattner. A common database approach for OLTP and OLAP using an in-memory column database. In *SIGMOD'09*, ACM (2009), 1–2.

[29] Kimmo E. E. Raatikainen. Cluster Analysis and Workload Classification. *SIGMETRICS Performance Evaluation Review* **20**(4) (1993), 24–30.

[30] Michael Stonebraker, Daniel J. Abadi, Adam Batkin, Xuedong Chen, Mitch Cherniack, Miguel Ferreira, Edmond Lau, Amerson Lin, Samuel Madden, Elizabeth J. O'Neil, Patrick E. O'Neil, Alex Rasin, Nga Tran, and Stanley B. Zdonik. C-Store: A column-oriented DBMS. In *VLDB'05*, VLDB Endowment (2005), 553–564.

[31] Ricardo Jorge Santos and Jorge Bernardino. Real-time data warehouse loading methodology. In *IDEAS'08* (2008), 49–58.

[32] Transaction Processing Performance Council. TPC BENCHMARK^{TM} H. White Paper, April 2010. Decision Support Standard Specification, Revision 2.11.0.

[33] Alejandro A. Vaisman, Alberto O. Mendelzon, Walter Ruaro, and Sergio G. Cymerman. Supporting dimension updates in an OLAP server. *Information Systems* **29**(2) (2004), 165–185.

[34] Youchan Zhu, Lei An, and Shuangxi Liu. Data Updating and Query in Real-Time Data Warehouse System. In *CSSE'08* (2008), 1295–1297.

[35] Marcin Zukowski, Niels Nes, and Peter Boncz. DSM vs. NSM: CPU performance tradeoffs in block-oriented query processing. In *Proceedings of the 4th international workshop on Data management on new hardware*, DaMoN'08, ACM (2008), 47–54.

[36] Daniel C. Zilio, Jun Rao, Sam Lightstone, Guy M. Lohman, Adam J. Storm, Christian Garcia-Arellano, and Scott Fadden. DB2 Design Advisor: Integrated automatic physical database design. In *VLDB'04*, VLDB Endowment (2004), 1087–1097.

[37] Daniel C. Zilio, Calisto Zuzarte, Sam Lightstone, Wenbin Ma, Guy M. Lohman, Roberta Cochrane, Hamid Pirahesh, Latha S. Colby, Jarek Gryz, Eric Alton, Dongming Liang, and Gary Valentin. Recommending materialized views and indexes with IBM DB2 Design Advisor. In *ICAC'04* (2004), 180–188.

Multiple-Site Distributed Spatial Query Optimization using Spatial Semijoins

Wendy OSBORN[a],[1] and Saad ZAAMOUT[a]
[a]*Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, Alberta T1K 3M4, Canada*

Abstract. In this paper, we present our strategy for distributed spatial query optimization that involves multiple sites. Previous work in the area of distributed spatial query processing and optimization focuses only on strategies for performing spatial joins and spatial semijoins, and distributed spatial queries that only involve two sites. We propose a strategy for optimizing a distributed spatial query using spatial semijoins that can involve any number of sites in a distributed spatial database. In this initial work we focus on minimizing the data transmission cost of a distributed spatial query by identifying and initiating semijoins from the smaller relations in order to reduce the larger relations and minimize the cost of data transmission. We compare the performance of our strategy against the naïve approach of shipping entire relations to the query site. We find that our strategy minimizes the data transmission cost in all cases, and significantly in specific situations.

Keywords. Spatial data, distributed spatial queries, optimization, performance, data transmission cost

Introduction

A distributed spatial database system [12] consists of several spatial database sites that are dispersed geographically. Each site manages its own collection of spatial data, but work collectively for processing inter-site data requirements. An important requirement of a distributed spatial database is the ability to efficiently process a query that requires spatial data from multiple sites. Historically, research in distributed relational databases focused on generating query execution plans that minimized the cost of data transmission over the network [3, 4, 11]. However, spatial data is more complex than alphanumeric data. This increases the complexity of joining spatial relations. Therefore, CPU and I/O costs should also be considered when processing a distributed spatial query [12].

Most existing strategies that process a distributed spatial query only work for two sites. The one exception to this does not handle spatial joins. Therefore, this preliminary work begins to address this shortcoming by propose a strategy for processing and optimizing a distributed spatial query that handles more than two sites. Our strategy focuses on minimizing the data transmission cost of a query by applying spatial semijoins in a cost-effective manner. An evaluation of our strategy shows

[1] Corresponding Author

reduction in the data transmission cost over the naïve approach (i.e. the approach that involves shipping entire relations directly to the query site). In specific situations, this reduction in cost is significant.

1. Background and Related Work

Most research in distributed spatial query processing focuses on spatial join algorithms, spatial semijoins algorithms, and the use of Bloom filters for processing distributed spatial queries. A spatial join [12] takes two relations R and S, each with a spatial attribute, and relates pairs of tuples between R and S based on a spatial predicate that is applied to the spatial attribute values. Examples of spatial predicates include the overlap, containment, and adjacency of two spatial attributes. A spatial semijoin [2] is performed by projecting the spatial attribute from one relation, transmitting it to the site that contains the other spatial relation, and performing a spatial join of the spatial projection and relation. Then, the qualifying tuples from second site are shipped back to the first site and joined with the spatial relation on that site. A Bloom filter [1] is a hashed bit array that provides a compact but imprecise representation of the values of a joining attribute. A '1' bit represents the possible existence of a joining attribute value, while a '0' bit represents the absence of the value.

With the exception of [6], all proposed strategies for processing distributed spatial queries work for a two-site database only. We summarize these works below.

1.1. Spatial Join

A significant majority of spatial join algorithms are designed for a centralized system [7]. This section focuses on spatial join strategies for distributed spatial queries.

Kang et al. [8] propose a parallel spatial join strategy that is adapted to a distributed spatial database environment. Their strategy has two phases: data redistribution, and filter and refinement. In the data redistribution phase, on each site, the space that contains objects is partitioned into regions (i. e. buckets). A subset of regions is transmitted between servers so that each server has the same regions for both data sets. This subset is chosen by estimating which will result in the lowest overall response time (although it is unclear if I/O costs are considered). Then, the filter and refinement phase is carried out on both sites by performing a spatial join. An experimental evaluation shows that the parallel spatial join technique has a significantly faster response time – up to a 33% improvement over a semijoin-based strategy.

1.2. Spatial Semijoin

Abel *et al.* [2, 13] propose a spatial semijoin operator that combines the conventional semijoin operator with the filter stage of spatial query processing in order to reduce the data transmission, I/O and CPU costs. Their work explores two adaptations of the spatial semijoin. In the first, a "projection" of a set of MBRs from one spatial relation is transmitted to the second site and applied to the other relation using a spatial join. In the second, the "projection" is a single-dimensional mapping that represents the objects in each relation. A performance evaluation between these approaches shows that 1) for datasets with very large spatial descriptions, both strategies perform the same, 2) for datasets with smaller spatial descriptions, a semijoin that uses single-dimensional mapping works best, 3) using the R-tree for retrieving MBRs incurs significant CPU costs, and 4) single-dimensional mapping causes more false drops than MBRs.

Karam and Petry [10] propose a spatial semijoin, which differs from [2, 13] in that MBRs from different levels of the R-tree are chosen for the spatial semijoin, instead of requiring that all come from the same level. A performance evaluation shows that their spatial semijoin outperforms the naïve spatial join (i. e. the whole relation is shipped to the other site for joining) when applied to real world data, but not when applied to randomly distributed rectangle sets. Limitation of their work are: 1) no comparison versus other strategies, 2) no consideration of CPU time.

1.3. Bloom Filters

Karam [9] propose a 2-dimensional bit-matrix approach for performing a semijoin of two relations that focuses on minimizing the data transmission, I/O and CPU costs. A 2-dimensional space is partitioned into equal-sized regions, with each region mapping to a bit in a 2-dimensional array. If a region contains objects, the corresponding bit is set to 1. This bit-matrix is transmitted to the site containing the other relation, and is applied by testing each region containing objects for the existence of a '1' bit in the bit-matrix. Any qualifying objects are sent to the first site. A performance evaluation shows that this approach shows the best improvement when applied to real world data. Limitation of this work are: 1) an evaluation against the spatial join, and not versus a spatial semijoin, which in functionality is a closer match to the bit-matrix approach, and 2) no compression of the bit-matrix – a bit-matrix that contains many zeros is still transmitted in its entirety.

Hua et al. [6] propose the BR-tree, which is an R-tree that is augmented with Bloom filters to support exact-match queries. Each node entry contains 1) a minimum bounding rectangle (MBR) that approximates an object or a subset of objects, and 2) a Bloom filter that also represents one or more objects. In a leaf node, a Bloom filter is created by taking each object and producing k bits in the filter using different hash functions. In a non-leaf node, a Bloom filter is created by intersecting the Bloom filters in its child node. Although the BR-tree supports exact-match queries by using Bloom filters, it still requires the MBRs for region and point queries. A strategy for processing distributed region, point, and exact-match queries is proposed. The algorithm duplicates the root of every BR-tree across every site in the database. Any objects that pass the test against a root node is shipped to the site containing the original BR-tree. This strategy works for any number of sites. A significant limitation is a lack of support for spatial joins.

2. Distributed Spatial Query Processing Strategy

In this section we present our algorithm for processing a distributed spatial query. The focus is to reduce the cost of data transmission over the network by using spatial semijoins. In the future we will also consider I/O and CPU costs.

2.1. Preliminaries

We apply spatial semijoins by shipping the smaller spatial attributes to other sites and applying them to the larger relations in order to eliminate a significant amount of data that will not participate in the final result. We utilize a modified version of the approximation based spatial semijoin that is proposed in [13].

In our implementation of the spatial semijoin, we use the following approach. We have two sites R and S that each contain a spatial relation. First, we obtain the projection of the spatial attribute from R. Then, the spatial relation is transferred to S and joined with the spatial relation on S. Our semijoins differs after this point. Instead of sending the qualifying tuples from S back to R for the final join, we send back to R the identifiers from the spatial projection, which are used to select tuples from the relation on R to ship to the final query site. In addition, the tuples on site S that qualified in the semijoin are also shipped to the query site. Using this semijoin strategy allows us to incorporate more than two sites when processing a distributed query.

We make the following assumptions in our work:

1. Every spatial object is represented using its minimum bounding rectangle (MBR).
2. Every site that participates in the distributed spatial query has one spatial relation. If a site contains other relations that are required for the query, it is assumed that all local processing has taken place and one spatial relations remains.
3. All spatial relations have one spatial attribute.
4. All objects (and corresponding MBRs) in all spatial attributes are drawn from the same "spatial domain" (i.e. same region of space).
5. The cardinality of each spatial attribute is equal to the number MBRs in the relation. That is, we assume that all MBRs in a spatial attribute are distinct.
6. The number of sites participating in the distributed spatial query is a multiple of two. The reason for this will be made clear when the algorithm is presented.
7. The spatial attribute for every spatial relation is already indexed by an R-tree (or a similar index that places the minimum bounding rectangles for all objects in its leaf level).

2.2. The Algorithm

Given n sites that will be participating in processing a distributed query, where each site has one spatial relation, our strategy has four main steps:

1. sorting and grouping by spatial attribute cardinality,
2. transmission of spatial attributes,
3. semijoin execution,
4. transmission of qualifying tuples to query site for the final join and processing.

Each step is described next. First, the sites will be ordered by increasing spatial attribute cardinality. After ordering, the first *n*/2 sites of the ordered list are placed in a set P, while the remaining *n*/2 sites are placed in a set Q.

Then, the spatial attribute from the relation on each site in P are transmitted to a site in Q in the following manner:

- the attribute from the site with the smallest spatial cardinality in P is sent to the site with the smallest spatial attribute in Q,
- the attribute from the site with the next smallest spatial cardinality in P is sent to the site with the next smallest spatial attribute in Q,
- and so on... until,
- the attribute from the site with the largest spatial cardinality in P is sent to the site with the largest spatial attribute in Q.

Next, on each site in Q, a spatial semijoin is performed between the existing spatial relation and the spatial attribute sent from the corresponding site in P. The results of of the semijoin are: 1) the set of tuples on the site that qualify in the semijoin, and 2) a set of identifiers from the spatial attribute whose MBRs also qualify in the semijoin. The set of identifiers is sent back to the corresponding site in P.

Finally, for all sites in P, the tuples whose identifiers match the ones obtained from Q are shipped to the query site. In addition, for each site in Q the set of tuples that qualified in the semijoin are sent to the query site. At the query site, the final join is performed.

2.2.1. Example

Suppose we have a distributed spatial database with six sites. Each site contains a spatial relation with 100, 200, 400, 600, 800, and 1000 tuples respectively. Our strategy for processing a query that involves these sites proceeds as follows. First, our sites are ordered by increasing spatial attribute cardinality. Then, the list is divided into the two sets. The set P will contain the sites with the 100-, 200- and 400-tuple relations, while the set Q will contain the sites with the 600-, 800-, and 1000-tuple relations.

Next, the spatial attributes from the sites in set P are sent to sites in Q in the following manner. First, the spatial attribute from the site containing 100 tuples is sent to the site that contains 600 tuples. Similarly, the spatial attribute from the 200-tuple sites is set to the 800-tuple site, and the spatial attribute from the 400-tuple site is sent to the 1000-tuple site. Then, on the 600-, 800-, and 1000-tuple sites, a semijoin is performed between the local spatial relation and the spatial attribute that was shipped to it. During this process, the identifiers that correspond to the MBRs in the spatial attribute that qualify for the semijoin are sent back to originating site. For example, on the 600-tuple site, the identifiers for the qualifying MBRs are sent back to the 100-tuple site, and are used to select the corresponding tuples. Finally, all qualifying tuples from all sites are shipped to the query site.

3. Experimental Evaluation

Here, we present our empirical evaluation of our distributed query processing algorithm. We compared our strategy for optimizing a distributed spatial query with the naïve approach which transfers all unreduced relations to the query site. First, we present the data sets and cost formulas used in our evaluation. Then, we present the results and discussion of our tests.

We simulated a six-site distributed spatial database, where each site contains one spatial relation. Each spatial relation has one spatial attribute, which consists of four values (*lx*, *ly*, *hx*, *hy*) that represent the extents of an MBR. In addition, each spatial relation has the following non-spatial attributes: identifier, region name, population and a line slope indicator. Each spatial relation has 100, 200, 400, 600, 800 and 1000 tuples respectively. We opted to use smaller relations for our experiments because of the preliminary nature of the work and the use of a simulated (and not real) distributed environment.

3.1. Data Transmission Cost Calculation

In our experiments, we estimated the cost of data transmission as the total number of bytes that are transmitted. We assume that the data transmission rate is constant and therefore is not added to our calculations. In addition, we assume a integer size of two bytes, a double-precision floating point size of eight bytes, a long integer size of 8 bytes and a character size of one byte.

The various costs of data transmission are calculated in the following manner. There are several calculations required. First, the transmission cost for transmitting an MBR is equal to the number of bytes used to represent an MBR:

$$cost\,(MBR) = 4 * sizeof\,(double) + sizeof\,(int) \tag{1}$$

which encompasses the co-ordinate values (*lx*, *ly*, *hx*, *hy*) and the tuple identifier. Similarly, the cost for transmitting a tuple is:

$$\begin{aligned} cost\,(tuple) = {} & sizeof\,(MBR) + 20 * sizeof\,(char) + sizeof\,(longint) \\ & + sizeof\,(int) \end{aligned} \tag{2}$$

which encompasses the region name, population and line slope indicator. In addition, 1) the cost of transmitting an identifier back to the original site from which it came is *cost* (*I D*) = *sizeof* (*int*), and, 2) the function *number_of_qualifiers* (*relation*) returns the number of tuples from a relation that participate in the result of a spatial semijoin operation.

Finally, given spatial attribute X from relation Y (i. e. site Y from the set P above) that is shipped to relation Z (i. e. site Z from set Q above), the cost of processing the spatial semijoin is:

$$\begin{aligned} cost\,(X, Y, Z) = {} & number\_of\_tuples\,(Y) * cost\,(MBR) \\ & + number\_of\_qualifiers\,(X) * (cost\,(ID) + cost\,(tuple)) \\ & + number\_of\_qualifiers\,(Z) * cost\,(tuple) \end{aligned} \tag{3}$$

The first term is the cost of transmitting the spatial attribute X from site Y to site Z. The second term is the cost of both transmitting back to Y the corresponding tuple identifiers for the qualifying MBRs in X, and then transmitting the tuples that correspond to those tuple identifiers to the query site. Finally, the third term is the cost of transmitting qualifying tuples from Z to the query site. This cost is calculated for every pair (Y, Z) of sites that are involved in the query, with all costs summed together to obtain the total cost of the query.

3.2. Two-Site Query Test

The first set of tests we performed are for distributed queries that involve two sites. Table 1 shows the pairs of relations (i.e. sites) that were evaluated, along with the total cost (in bytes) of both our optimized strategy (column Optimized) and the naïve approach (column Naïve). We opted to report the cost of data transmission in bytes so that determining the number of physical disk blocks in a page of secondary storage would not be required at this point.

In all cases, our strategy results in a lower data transmission cost over the naïve approach. In particular, the most significant improvement is achieved when there exists a significant difference in the size of the spatial relations between the two sites. For example, when the query involves the sites that contain the 100- and 1000-tuple spatial relations, we have almost 80% less data that is being transmitted when our strategy is being used to process the query.

Table 1. Two-site query test

Site 1	Site 2	Optimized	Naïve	%Improvement
100	400	16010	32000	50
100	600	16270	44800	64
100	800	15750	57600	73
100	1000	14580	70400	79
200	400	32150	38400	17
200	600	31760	51200	38
200	800	32020	64000	50
200	1000	31890	76800	59

Table 2. Four-site query test

Site 1	Site 2	Site 3	Site 4	Optimized	Naïve	%Improvement
100	200	400	600	52264	83200	37
100	200	800	1000	53410	134400	60
400	600	800	1000	162604	172900	6

3.3. Four-Site Query Test

For our second set of tests, we compared the evaluation of the strategies for four-site queries. Table 2 shows the sites involved and the total costs in bytes from both strategies. Again, we find that our strategy outperforms the naïve approach. In addition, we also find that in the situation where a significant size difference exists between the relations – in this case, 100, 200, 800, and 1000 tuples – the greatest improvement is achieved.

3.4. Six-Site Query Test

Finally, we performed one test that compares our strategy with the naïve approach when all six sites are involved. We found the transmission cost from the query optimization strategy to be 127,456 bytes and that from the naïve strategy to be 198,400. This gives an improvement of approximately 36%.

3.5. Discussion

In all cases, we discovered a lower data transmission cost from our strategy over the naïve approach. In addition, we discovered the following trends. First, the queries with the largest difference in the number of tuples between the participating relations, the greater the reduction that our strategy achieves. Second, we discovered that as the difference in the number of tuples between participating relations increases, the improvement that our strategy achieves increases as well.

4. Conclusion and Future Work

In this paper, we propose a strategy for optimizing queries in a distributed spatial database that involves relations on multiple sites. Our strategy focuses on minimizing the cost of data transmission by applying spatial semijoins. Smaller spatial attributes are chosen for transmission and application to larger relations so that overall data transmission costs are reduced. A empirical evaluation of our strategy against the naïve approach shows that our strategy achieves a reduction in the data transmission cost in all cases. In particular, as the size difference between relations increases, the savings achieved by our strategy over the naïve strategy are very significant.

As mentioned, one important direction of future work that we are currently exploring is the resulting I/O and CPU costs from our optimization strategy. It is important to determine if the I/O and CPU costs are minimal or outweigh any benefits of our strategy.

Other directions of future work include the following. One is to create a real distributed database system with multiple sites, which will provides a means for better evaluation of our strategy. Another is to evaluate the two-site version of our strategy (i. e. when only two sites are involved) versus other existing strategies. Although the focus of this work was to extend the number of sites involved in a distributed spatial query, evaluating the efficiency of our algorithm in the two-site case versus existing strategies is also important and would better identify if our strategy is superior in this situation. A final research direction is to develop and evaluate other strategy for processing and optimizing a distributed spatial query. As discussed, very limited work has been proposed, which leads to many exciting opportunities for research in the area of distributed spatial query processing.

References

[1] B.H. Bloom, Space/time trade-offs in hash coding with allowable errors, *Communications of the ACM* **13** (1970), 422–426.

[2] D.J. Abel, B.C. Ooi, K.-L. Tan, R. Power and J.X. Yu, Spatial join strategies in distributed spatial DBMS. In: *Proceedings of the 4th International Symposium on Advances in Spatial Databases*, 1995.

[3] P.M.G. Apers, A.R. Hevner and S.B. Yao, Optimization algorithms for distributed queries, *IEEE Transactions on Software Engineering* **9** (1983), 57–68.

[4] P. Bodorik, J.S. Riordon and J.S. Pyra, Deciding to correct distributed query processing, *IEEE Transactions on Knowledge and Data Engineering* **4** (1992), 348–357.

[5] T. Brinkhoff, H.-P. Kriegel and B. Seeger, Efficient processing of spatial joins using R-trees. In: *Proceedings of the 1993 ACM Sigmod International Conference on Managment of Data*, New York, USA, 237–246, 1993.

[6] Y. Hua, B. Xiao and J. Wang, BR-Tree: a scalable prototype for supporting multiple queries of multidimensional data, *IEEE Transactions on Computers* **58** (2009), 1585–1597.

[7] E. Jacox and H. Samet, Spatial join techniques, *ACM Transactions on Database Systems* **34** (2007), 1–44.

[8] M.-S. Kang, S.-K. Ko, K. Koh and Y.-C. Choy, A parallel spatial join algorithm for distributed spatial databases. In: *Proceedings of the 5th International Conference on Flexible Query Answering Systems*, 212–225, 2002.

[9] O. Karam, *Optimizing Distributed Spatial Joins using R-trees*, Ph.D. Thesis, Tulane University, 2001.

[10] O. Karam and F. Petry, Optimizing distributed spatial joins using R-trees. In: *Proceedings of the 43rd ACM Southeast Regional Conference*, 2006, 222-226.

[11] M.T. Özsu and P. Valduriez, *Principles of Distributed Database Systems*, Springer, New York, 2011.

[12] S. Shekhar and S. Chawla, *Spatial Databases: a Tour*, Prentice Hall, New Jersey, 2003.

[13] K.-L. Tan, B.C. Ooi and D.J. Abel, Exploiting spatial indexes for semijoin-based join processing in distributed spatial databases, *IEEE Transactions on Knowledge and Data Engineering* **12** (2000), 920–937.

Cost Models for Approximate Query Evaluation Algorithms

Oxana DOLMATOVA, Anna YARYGINA and Boris NOVIKOV
Saint Petersburg State University
oxana.dolmatova@gmail.com, anya_safonova@mail.ru, borisnov@acm.org

Abstract. The optimization is essential for any high-performance querying system. Several optimization techniques were developed and successfully implemented for relational databases. However, these techniques should be re-examined and revised for distributed heterogeneous systems of information resources supporting diverse querying paradigms. We introduce cost models for approximate query evaluation in the context of generalized algebraic operations supporting both exact and similarity queries. The proposed cost models are suitable for approximate evaluation and trade-off between computational performance and the quality of results. We present a rationale for our approach and elaborate our cost model for key operations and algorithms.

Keywords. Cost models, approximate algorithms, query evaluation, heterogeneous systems, information resources

Introduction

The presence of high-level declarative query languages was considered as one of major strengths of database management systems since early 70-ies and became an inherent feature of the relational database model and its variations are implemented in the industrial SQL-based systems.

Providing highly expressive declarative means for query specification, these languages both require and enable sophisticated optimization. In contrast with low-level imperative object-oriented programming languages, which allow only moderate code improvements with local optimization, declarative languages depend dramatically on the quality of the optimizer.

Formally, the task of a query optimizer is to choose an algebraic expression of minimal execution cost among several equivalent expressions. In other words, any high-quality optimizer is inevitably a cost-based one and, hence, the cost model is one of the critical core components of the optimizer.

The abstract concept of cost may include several different measures of query execution performance. Usually the most important is execution time (either CPU or elapsed), amount of I/O, or, in a distributed mobile environment, the battery energy. However, the actual interpretation of cost is not essential for the optimization process.

There are many optimization techniques based on cost models, but in general almost all cost models were prepared for exact query evaluation algorithms.

In broad modern contexts, such as distributed heterogeneous systems, real-time business analytics and distributed mobile environments an approximate query evaluation becomes a must. Indeed, it does not make any sense to use exact query evaluation in uncertain context or for similarity-based queries, where "top k" approach is the best. Approximate query evaluation is ultimately needed to provide timeliness for business analytics or save energy in a mobile device.

As the query evaluation is approximate, the quality of the output may decrease. Our main objective is to provide a cost model capable to support trade-off between the cost of evaluation and the quality of results. Based on our cost model, the query optimizer can either provide best possible quality for a given cost, or minimize the cost providing at least required quality of the query output.

We start from naïve exact cost models and proceed with detailed analysis of approximate operations. We define models for a number of operations including "top k" operations for single and multiple arguments.

In this research we consider query processing in a heterogeneous environment of autonomous information resources. In this type of environments data statistics might be unavailable, hence simple cost models are needed for both of exact and approximate querying.

1. Cost Model Specification

In this paper we discuss query processing in a heterogeneous environment of autonomous information resources where all objects have scores which represent their relevance, similarity, quality and so on. Each operation takes as an input a stream of objects with scores and pushes into the output another stream of objects with their scores.

Further the specification of the cost model that supports the concept of operation result quality is presented and is followed by brief discussion of main operations and corresponding algorithms which are analyzed in this paper.

Operation cost model depends on 4 basic parameters: data size; data quality; data order; and the amount of resources needed for operation execution (operation cost).

On the one hand in traditional cost models the operation cost (amount of resources for its evaluation) can be estimated based on the size of arguments, order of objects in an argument. On the other hand, when we talk about approximate algorithms, the operation behavior depends on the quality of the arguments, the size of the arguments, order of objects in the arguments, and the amount of resources allocated for operation execution.

Thus, an operation cost model is defined as follows:

$opcost(resources,\ size_{in,}order_{in})=(quality_{out},size_{out,}order_{out}),$ where

- *opcost* is the operation cost function, *opcost* connects the values of the input and output parameters;
- *resources* are resources allocated for the operation processing or the operation cost;
- $size_{in}$ is the operation input arguments size;
- $order_{in}$ is the operation input arguments order;
- $quality_{out}$ is the operation execution relative result quality, the relative quality of the result of the corresponding subquery;

- $size_{out}$ the result size;
- $order_{out}$ the result order;
- $resources$, $size_{in,}$ $order_{in}$, $quality_{out}$, $size_{out,}$ $order_{out}$ are objects with attributes which represent different measures of the corresponding parameter of the cost model and are discussed in details below.

Let us note that $size_{in}$ and $order_{in}$ are vectors of objects that describe all arguments of the operation.

The equation describes the cost model and binds all its variables and parameters. Thus, substituting in the equation the values of the variables, we can express and evaluate the remaining parameters of the cost model.

Apart from the basic cost model equations there are some restrictions on the cost models of the specific operations:

$$resources^{min} \leq \text{resources} \leq resources^{max}$$
$$quality^{min} \leq \text{quality} \leq quality^{max}$$

If the operation has $resources^{max}$ resources available, the result has the best possible quality with given arguments, that is relative $quality_{out}=quality^{max}$.

If the operation has $resources^{min}$ resources available, the result has the worst possible relative quality ($quality_{out}=quality^{min}$) with given arguments. In this case the algorithm spends the least possible amount of resources for operation execution.

Our cost models suggest the estimation of the unknown parameters. We assume that objects are not ordered in an input or output data if no information about order is available. The value of the quality and size parameters are estimated based on the history of the previous queries and collected statistics.

Further we discuss how to measure resources and data characteristics in our cost models.

1.1. Resources

Resources needed for a particular operation processing can be measured and evaluated in different ways. For example the operation processing cost can be estimated by its execution time. However, this approach restricts the proposed cost model to a specific query processing system.

In this research we evaluate the cost of operations in terms of the number of disk accesses, sequential accesses, executions of external operations, and so on. It is important to note that in this case, the cost model is more general. In this case the proposed cost model can be tuned depending on the specific characteristics of the query processing engine.

There are several characteristics and measure resources:

- *sa* the number sequential accesses to read objects from pipe;
- *dsa* the number of sequential disk accesses to read objects;
- *ma* the number of random memory accesses;
- *da* the number of random disk accesses;
- *pred* the number of predicate evaluations;
- *ha* the number of hash table accesses.

The operation cost or the amount of resources needed for its execution will be evaluated in terms of these measures. The cost of each parameter in terms of time ($T_{sa,}$

T_{dsa}, T_{ma}, T_{da}, T_{pred}), which depends on the system configuration, will be estimated based on the experiments and available statistics. Thus the operation cost can be simply evaluated:

$$time = sa*T_{sa}+dsa*T_{dsa}+ma*T_{ma}+da*T_{da}+pred*T_{pred}+ha*T_{ha}$$

1.2. Size

The size of the input data, as well as the result sizes can be evaluated at least in two different ways: size, which data occupy in memory (*size*), and the number of objects, semantic units, which the operation processes (*cardinality*). These characteristics influence the behavior of algorithms that implement operations.

1.3. Quality

We distinguish absolute and relative quality which operations produce.

The absolute quality shows how the produced result suites to the user expectations. To measure the absolute quality we need to obtain the actual relevance scores which are usually not know in advance and sometimes at all.

The relative quality shows how the limited, approximate implementation of an operation changes the absolute quality of the result. In our cost models we operate with relative quality of the operations and queries. The most important property of the relative quality of a query is its monotony on the amount of resources allocated for its processing. The monotony of the relative quality depends on its definition and algorithms implementing approximate operations.

1.4. Order

The order of objects in the input data influences the operation execution cost. At this phase of work, the order of objects is defined as the order of their scores in a data set. The cost models depend on the fact whether the objects in the input data are naturally ordered according to their scores.

2. Operations and Algorithms

We consider three operations in this paper. Exact and approximate algorithms implementing them are discussed in this section and corresponding cost models are developed and analyzed further.

2.1. Top k

Top k operation takes as an input the stream of objects with scores and returns to the output the stream of best k objects according to their input scores. The naïve exact algorithm orders objects according to their scores if needed and returns k first of them. An approximate top k algorithm reads objects sequentially, while free time for operation execution is available, and pastes them into the sorted list of already processed objects. When the resources are over first k objects are set to be the result.

2.2. Fusion

Binary fusion operation processes two input streams of objects with scores and returns stream of received objects with newly generated aggregated scores. The naïve exact algorithm implementing fusion is based on nested loops.

2.3. Aggregation

The aggregation operation is the same as defined in [7]. The operation takes as an input two streams, where objects are naturally ordered according to their scores, and returns to the output best k objects according to the aggregated scores based on the input ones. The naïve exact algorithm is a simple combination of fusion and top k operation. However, effective and optimal algorithms for aggregation operation were developed in [7]: FA (Fagin's Algorithm) and NRA (No Random Access Algorithm). Because the input streams are coming from two independent sources and random accesses in the first algorithm can be expensive and sometimes impossible, they have been removed from the implementation of algorithms and replaced with sorted accesses.

3. Cost Models for Exact Algorithms

Let's describe a basic cost models for different operations in case when the relative quality of the arguments and the result of the operation as the best possible and is not regulated from the outside.

Actually cost models depend on algorithms implementing this operations rather than operations themselves. Here we describe cost models for natural exact operation algorithms discussed in section 3.

All restrictions will be defined based on the cost model described by the equation:

$opcost(resources, size_{in,} order_{in}) = (quality_{out}, size_{out,} order_{out})$.

3.1. Top k

For exact top k operation the cost model can be defined as follows:

```
$cardinality_{out}=min\{k, cardinality_{in}\}$
$order_{out}=true$
$cardinality_{in}\geq cardinality_{out}$
$size_{in}\geq size_{out}$
if $order_{in}=false$ then
    $sa=cardinality_{in}$
    $ma=C*cardinality_{in}*ln(k)$
if $order_{in}=true$ then
    $sa=cardinality_{out}$
```

3.2. Fusion

Here we consider the binary fusion operation. In this case $size_{in,}order_{in}$ represent two-dimensional vectors, since the fusion operation takes two arguments. For fusion operation (based on nested loop algorithm) we have the following cost model:

$sa=\sum cardinality_{in}$
$dsa=\sum cardinality_{in}$
$da=\prod cardinality_{in}$
$cardinality_{out}\leq\sum cardinality_{in}$

3.3. Aggregation

First of all we describe the intermediary parameters which can help us to construct formulas:

- t – estimated size of a table, which shows all objects with already read scores;
- ns – estimated number of scores needed to fill the table of size t;
- ks – estimated number of scores needed to fill the table of size k.

We assume independence of scores and uniform distribution and use expectation for evaluate these parameters.

$order_{in}=true$
$cardinality_{out}=min\{k, cardinality_{in}\}$
$order_{out}=true$
$cardinality_{in}\geq cardinality_{out}$
$size_{in}\geq size_{out}$

3.3.1. FA

$time = cardinality_{in}T_{sa}+(t\log t+4t)T_{ma}+tT_{ha}$

Summands include sorted access, sort, input/output and calculation of aggregate score as well as search for a place to insert a score just obtained from stream respectively.

After evaluation and substitution of intermediary parameters we obtain the following:

$time=((3/4\ln(cardinality_{in})+3/2)T_{ma}+3/4T_{ha})k+T_{ma}cardinality_{in}+1/4cardinality_{in} T_{ha}+1/4cardinality_{in}(ln(cardinality_{in})–2)T_{ma}+cardinality_{in}T_{sa}$

3.3.2. NRA

$time=cardinality_{in}T_{sa}+((ns-ks)/2t\log t+2t(2+t))T_{ma}+tT_{ha}$

Summands include sorted access, sort, input/output and calculation of worst case score and best case score as well as search for a place to insert a score just obtained from stream respectively.

After evaluation and substitution of intermediary parameters we obtain the following (c_{in} denotes as $cardinality_{in}$ below):

$time=(45/32-9/64\ln(c_{in}))T_{ma}k^2+((3/4c_{in}+3-3/64c_{in}(\ln(c_{in})–2)+3/16c_{in}(3/4\ln(c_{in})– 3/2)T_{ma}+3/4T_{ha})k+(1/2c_{in}(2+1/4c_{in})+3/64c_{in}{}^2(\ln(c_{in})–2)T_{ma}+c_{in}T_{sa}+1/4c_{in}T_{ha}$

4. Cost Models for Approximate Algorithms

4.1. Aggregation

Now we add the specific resource t0 – time operation execution. The problem is to estimate the result quality.

We create approximate model based on exact one. Let's *k'* denote the number of correct object returned by approximate algorithm. We express it in terms of *t0*. Hence we get number of scores which system can find for the given time. Then we consider ratio between *k'* and *k* which means the accuracy of result.

$quality_{out}=k'/k$

For FA we obtain the following: $k'=(t0-T_{ma}cardinality_{in}-1/4cardinality_{in}T_{ha}-1/4cardinality_{in}(\ln(cardinality_{in})-2)T_{ma}-cardinality_{in}T_{sa})/((3/4\ln(cardinality_{in})+3/2)T_{ma}+3/4T_{ha})$.

We will not give the inverse formula for NRA here because of the limits of paper size.

4.2. Top k

For approximate algorithm implementing top k operation we have:

$cardinality_{out}=\min\{k,cardinality_{in}\}$
$order_{out}=\text{true}$
$cardinality_{in}\geq cardinality_{out}$
$size_{in}\geq size_{out}$

Let us estimate $size_{min}<size_{out}<size_{max}$, $resources_{min}<resources_{out}<resources_{max}$, $quality_{out}^{min}<quality_{out}<quality_{out}^{max}$ of the operation processing result:

$cardinality_{min}=cardinality_{out}=cardinality_{max}=\min\{k,cardinality_{in}\}$,
$size_{min}=size_{out}=size_{max}$,
if $order_{in}=\text{true}$ then
  $sa_{min}=sa=sa_{max}=cardinality_{out}$
  $quality_{out}^{min}=quality_{out}=quality_{out}^{max}=1$
if $order_{in}=\text{false}$ then
  $sa_{min}=cardinality_{out}$
  $sa_{max}=cardinality_{in}$
  $sa=S$
  $ma_{min}=0$
  $ma_{max}=C*cardinality_{in}*\ln(k)$
  $ma=C*S*\ln(k)$
  $quality_{out}^{max}=1$
  $quality_{out}^{min}=0$
  $quality_{out}=S/cardinality_{in}$ (in case when object scores are distributed uniform)

5. Related Work

The query optimization became both required and enabled since the advent of high-level declarative query languages, mostly in the context of the relational database model.

A brief overview of classical query optimization techniques can be found in [11]. The optimization techniques for distributed systems are summarized in [10]. An optimizer for distributed heterogeneous systems is proposed in [15].

The cost models proposed in this research are designed similar to those needed to traditional optimizers.

The optimization strategy based on algebraic equivalences between similarity based operations that serve as rewrite rules is outlined in [4]. Optimization rules based on similarity based algebraic framework properties and equivalence laws are also discussed in [1, 14, 5, 12].

It is important to mention that the lack of algebraic equivalences pushes the research to the development of optimization techniques based on performance/quality tradeoffs and approximate algorithms.

The approximate query evaluation techniques were considered in the context of very large data warehouses and mobile networks [2, 6, 3, 8]. The approximation is typically based on sampling, wavelets, or synopsis.

Handling of time constraints on complex SQL queries is proposed in [7]. The authors distinguish approximate (based on sampling) and partial (top k) query evaluation.

The quality and performance trade-offs for stream processing are discussed in [16, 9].

Cost models based on estimation of operation selectivity and cardinality are introduced in [1] for the selected set of operations: union, intersection, and difference; joins; merge; subtract; select; and map. Optimization rules based on the query tree reconstruction are derived from the previous analysis.

A sampling-based method to estimate the cardinality of rank-aware operators is developed for costing plans [12].

A novel multi-criteria query optimization techniques for performing query optimization in databases, such as multimedia and web databases, which rely on imperfect access mechanisms and top-k predicates are proposed in [13]. The size and quality factors are introduced into the cost model and optimization algorithms.

6. Conclusion

In this paper we presented several cost models for both exact and approximate query evaluation algorithms in distributed heterogeneous systems. The more resource we use, the more accurate the result is and vice versa. Our models convert these intuitive observations into clear and distinct formulas. It provides formalism for trade-off between quality and performance.

In the future we are going to conduct experiments in order to estimate the accuracy of our cost models.

References

[1] S. Adali, P. Bonatti, M. L. Sapino, and V. S. Subrahmanian, A multi-similarity algebra. In: *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD '98)*, New York, NY, USA. ACM, 1998, 402-413.

[2] B. Babcock, S. Chaudhuri, and G. Das, Dynamic sample selection for approximate query processing. In: *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD'03),* New York, NY, USA. ACM, 2003, 539-550.

[3] S. Chaudhuri, G. Das, and V. Narasayya, Optimized stratified sampling for approximate query processing, *ACM Transactions on Database Systems* **32**(2) (2007), 9.

[4] S. Chaudhuri, R. Ramakrishnan, and G. Weikum, Integrating DB and IR technologies: what is the sound of one hand clapping? In: *Proceedings of Second Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA, January 4-7, 2005, 1-12.

[5] P. Ciaccia, D. Montesi, W. Penzo, and A. Trombetta, Imprecision and user preferences in multimedia queries: a generic algebraic approach. In: *Foundations of Information and Knowledge Systems, Proceedings of First International Symposium (FoIKS'2000)*, Burg, Germany, February 14-17, 2000. Lecture Notes in Computer Science **1762** (2000), Springer, Berlin, 50-71.

[6] C. Dell'Aquila, F. Di Tria, E. Lefons, and F. Tangorra, Accuracy estimation in approximate query processing. In: *Proceedings of the 14th WSEAS International Conference on Computers: Part of the 14th WSEAS CSCC Multi-Conference (ICCOMP'10)* **1,** Stevens Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS), 452-458.

[7] R. Fagin, A. Lotem, and M. Naor, Optimal aggregation algorithms for middleware, *Journal of Computer and System Sciences* **66**(4) (2003), 614-656.

[8] Y. Hu, S. Sundara, and J. Srinivasan, Supporting time-constrained SQL queries in Oracle. In: *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)*, VLDB Endowment, 2007, 1207-1218.

[9] Ch. Jermaine, S. Arumugam, A. Pol, and A. Dobra, Scalable approximate query processing with the DBO engine, *ACM Transactions on Database Systems* **33**(4) (2008), 23:1-23:54.

[10] Q. Jiang, *A Framework for Supporting Quality of Service Requirements in a Data Stream Management System*. PhD thesis, Arlington, TX, USA, 2005.

[11] D. Kossmann, The state of the art in distributed query processing, *ACM Computer Survey* **32**(4) (2000), 422-469.

[12] D. Kossmann and K. Stocker, Iterative dynamic programming: a new class of query optimization algorithms, *ACM Transactions on Database Systems* **25**(1) (2000), 43-82.

[13] Ch. Li, Kevin Ch.-Ch. Chang, I. F. Ilyas, and S. Song, RankSQL: query algebra and optimization for relational top-k queries. In: F. Ozcan, editor, *SIGMOD Conference*. ACM, 2005, 131-142.

[14] L. P. Mahalingam and K. S. Candan, Multi-Criteria Query Optimization in the Presence of Result Size and Quality Tradeoffs, *Multimedia Tools and Applications Journal* **23**(3) (2004), 167-183.

[15] D. Montesi, A. Trombetta, and P. A. Dearnley, A similarity based relational algebra for web and multimedia data, *Information Processing and Management* **39**(2) (2003), 307-322.

[16] F. Pentaris and Y. Ioannidis, Query optimization in distributed networks of autonomous database systems. *ACM Transactions on Database Systems* **31**(2) (2006), 537-583.

[17] R. Zhang, N. Koudas, B. Ch. Ooi, D. Srivastava, and P. Zhou, Streaming multiple aggregations using phantoms, *The VLDB Journal* **19**(4) (2010), 557-583.

Processing Multiple Databases in the Estonian Water Information System

Vladimir VIIES[a], Peeter ENNET[b], Jaan AIGRO[a], Hannes KINKS[a], Robert KULLAMAA[a], Ott Madis OZOLIT[a] and Ain SALULA[a]

[a] *Tallinn University of Technology, Estonia*

[b] *Estonian Environmental Information Centre, Estonia*

Abstract. For the purpose of gathering and publishing Open Data, the Estonian Environmental Information Center (EEIC) has numerous databases of different architecture and structure available. An Estonian Water Information System (EWIS) is being planned and already being developed using these databases. Countries such as the USA and UK have already taken leaps to have their it available to the general public, and Estonia is following suit. In line with the Aarhus Convention (Aarhus, 1998), which states that environmental data, which is also considered to be Open Data, needs to be readily available to the general public. The main issues appearing with the data is that due to the structural differences, it is difficult to make the data contained in them interoperable in the EWIS. This issue will be tackled by a database interface, which will be designed for using the EEIC databases in co-operation, formatting the data on demand to a form which the EWIS will easily recognize and utilize within its applications. The information system itself will be used for providing public services, including simple queries about water condition, specific queries for water parameters, overview of the Estonian waters and modeling data to predict outcomes to an user-defined situation. Queries will be responded to in real-time, including those which need models to process data before returning it to the user. This means the EWIS will be a tool to be used widely, not limiting itself to Estonians, as foreigners might take an interest in the condition of Estonian waters as well. This also makes the EWIS a very suitable tool for environmental specialists from any country, such as hydrologists, to scrutinize Estonian watersheds, for example. In the future, it is a possibility that the EWIS will be integrated within the Estonian national data exchange grid, X-road (X-tee in Estonian), as a service. General information and appraisal will be offered, for people who are more interested in data on more of a black and white scale, meaning the system will estimate if something about a river or lake is either bad or good. In the future, the EWIS will be designed to offer these services from a cloud-based platform. These services will at the same time be offering Open Data to its users, as all environmental information is defined as such. Any sort of data processing applied by the EWIS will also be considered as Open Data, and no monetary compensation for the processing shall ever be asked from the end-user.

Keywords. Cloud services, database compatibility, data modeling, open data, water quality

Introduction

For various reasons, a number of different databases for storage of environmental data have been created in Estonia. As a rule, these databases are not connected to or dependent on each other and, therefore, obtaining generalized data may cause hardship. Additionally, all environmental data is also considered to be Open Data, according to the Aarhus Convention [1] and and Open Data principles, both being actively put applied by the Estonian government [2]. The Estonian Water Information System (EWIS) has to obtain data mainly from three EEIC databases – the Environmental Register, the IS of Environmental Permits and the IS of Estonian Nature. In addition, there is need to query data from several other databases managed by other agencies or institutions. One disadvantage, which appears in the design of sectorial databases in Estonia is, that they are made under administrative guidelines to achieve only some specific tasks. There are cases where significant information is stored in the database as, for example, comment fields, because for that particular task, this information was not essential, but considered relevant enough to store in some simple form. Such information is difficult to access and process. One source of the problems is that different databases are not using the same standards. An evident example is the usage of units. Despite the fact that SI-system units are obligatory for historical reasons, different units are used. For example, the dissolved oxygen concentration in marine data is presented as millimoles per liter while in river monitoring it is presented as milligrams per liter. Briefly said, there are many difficulties to be resolved by using different databases in a single, large geographical information system (GIS). When more developed, we wish EWIS to appear as a public service in the Estonian national portal, linked through the national data exchange grid, X-tee (X-road) [3].

1. Data Infrastructure

One principle that we have in the project is that we use and process Open Data, a term coined by Tim Berners-Lee. Through said processing, we generate new Open Data, also freely available for general use [4]. This Open Data will be made available to the general public through a map-based application, which is designed mainly according to the needs of environmental specialists, yet useable by anyone with interest towards hydrological and hydrochemical data or the environment in general.

As we are dealing with the issue of handling many databases which store similar, albeit differently formatted data, we need to have a link between our applications and data sources which clears the discrepancies. The goal is to make the databases work in unison through a database interface, nicknamed ROOMA [5], as shown in Figure 1. It will act as a translator between the different databases when a multiple-target query is made so that the result is always machine-readable and processable in the mathematical models (or applications) that we use.

We decided to combat the issue of processing aforementioned databases using a mediated approach. ROOMA is used for all databases, which means that bloating can become a problem, as a system that wishes to do everything may crumble under its own weight. Therefore we decided to be lightweight in the realization of the system.

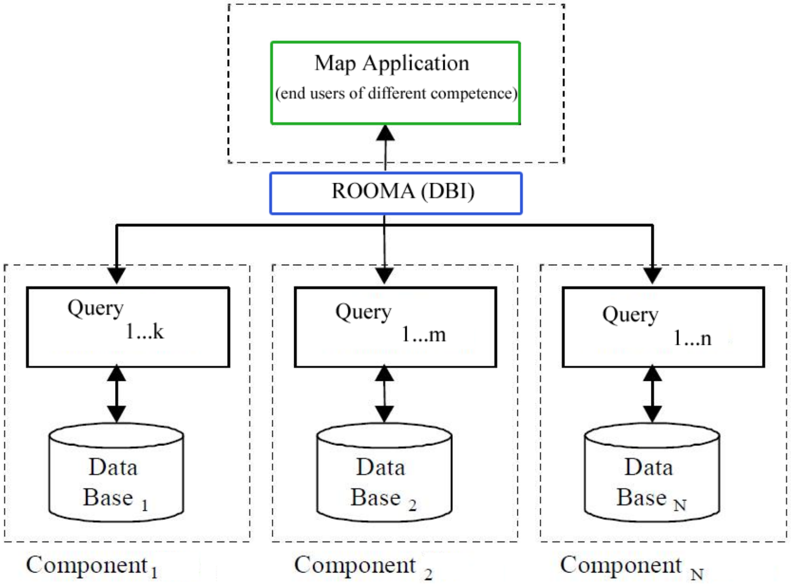


Figure 1. Data integration schematic of the EWIS

For this, we use Java as the language of choice, alongside the Hibernate library. Since we are also dealing with databases that run on different engines, Hibernate allows to process queries in object form, minimizing the needs for writing large amounts of SQL code. ROOMA will also deal with the formatting of similar data to a uniform format so that applications will get information that is converted to their used measurement systems as required.

The purpose of ROOMA is to offer open, structured, homogenized information and methods of delivering it, such as Plain Old Java Objects (POJOs), which we are using right now, and XML in the future. We also want to keep data retrieval as simple as possible for developing the system. To homogenize the data we are dealing with, ROOMA does not offer specific applications with specific queries connected as binary relations – the idea is to map the information in EWIS. Figure 2 shows EWIS' structure on a class level. Each *RequestObject*, which are pre-made, contains the information the application or client needs, ROOMA then decides upon which data to query based on the *RequestObject*. It queries all the relevant databases for information and returns a *ResultObject* to the client application. The clients themselves do not hold any connection information for any database, it is purely handled by ROOMA, to reduce security risks. *RequestObjects* also have restriction-expansion objects, built using the *Builder pattern*, related to them (as a filter) to avoid under- or over-querying data. For example, if a client requests information about a river, and does not have a restriction-expansion object related to it, then ROOMA will understand that it needs to query data only from the *River* class, and not from any sub-class related to it, e.g. the *RiverMeasurement* class.

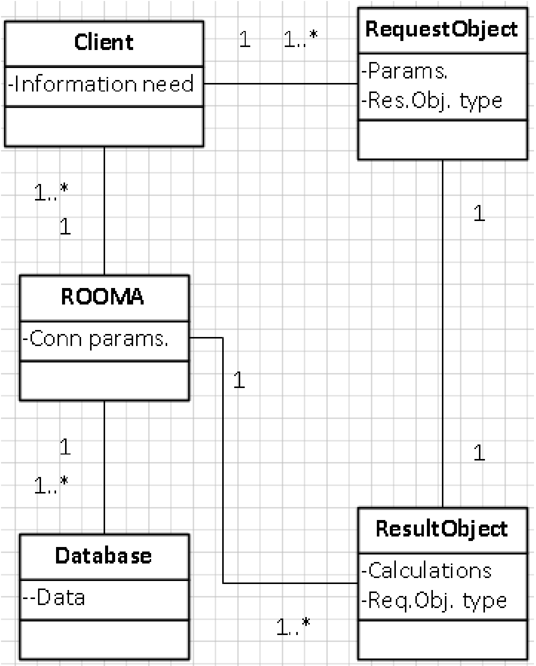


Figure 2. EWIS object structure

Each query is comprised of three main objects – Request, Handler and Result. The Request and Result objects are visible and usable by all components, while the Handler is strictly used in ROOMA for security reasons. When ROOMA receives a request, it sends it to a *HandlerObject* which does the actual queries to our databases, the *HandlerObject* then builds a *RequestObject* and it is sent through ROOMA back to the client. ROOMA works on asynchronous principles – we use the Java Netty framework for object handling in ROOMA, as well as the communication platform for all our Java solutions. Hibernate allows us to do queries in both SQL and object form (which is still converted to SQL since we are dealing with SQL-family databases). The query flow is as follows:

1. *RequestObject* is created.
2. A filter/wrapper is added to the *RequestObject.*
3. *RequestObject* sent to ROOMA.
4. ROOMA provides the corresponding handler the *RequestObject.*
5. The *HandlerObject* performs the queries from our databases.
6. When/If ready, a *ResultObject* is created and dispatched back to the client.

As for data storage, the EWIS aspires towards cloud computing. This is desirable because of two main aspects: data pricing and mobility. There are already agreements with Swedbank Estonia to provide resources and knowledge to transition at least some of our data sources into a cloud-based system. The software developed in our system will be thus offered as Software as a Service (SaaS) when the transition is made.

2. Information System Design

When designing the EWIS, we have to approach from two very different standpoints: the perspective of the user and the perspective of the solution. Designing the EWIS, we have to base it on the needs of environmental specialists first and foremost. Their job demands the following aspects of the information system:

- Map-based interface.
- Web-based applications.
- Ability to form environmental reports.

If we take these requirements into account, it is not hard to make the system useable to the general public as well, not just environmental specialists. Ease of use will inevitably be a key element in the design, since said specialists might not be computer savvy. Each component (or application) should be useable by a professional to help do their jobs, an enthusiast to satisfy the need for information and also teaching personnel, in environmental courses for example.

When taking measurements and monitoring data, we have to consider the magnitude of the data involved, as shown in Figure 3. Some detailed models require a lot of data as an input and give almost as much output data. Data sizes are described as amount of data per year. For example: *In-situ* measurement data reaches sizes to a few gigabytes. When using indirect measuring, the data size is multiplied hundredfold – a few hundred gigabytes. And last but not least, the data sizes for modeling can reach a few terabytes per year. Another advantage of cloud computing that could be applied to our information system is cloud storage. The data can be stored in remote, virtual databases, still maintaining constant access to the data, all while cutting data pricing. To counter this issue, the information system will, at first, provide seasonal results as a maximum modeling precision, since real-time prognosis would require constant, immeasurable data flow.

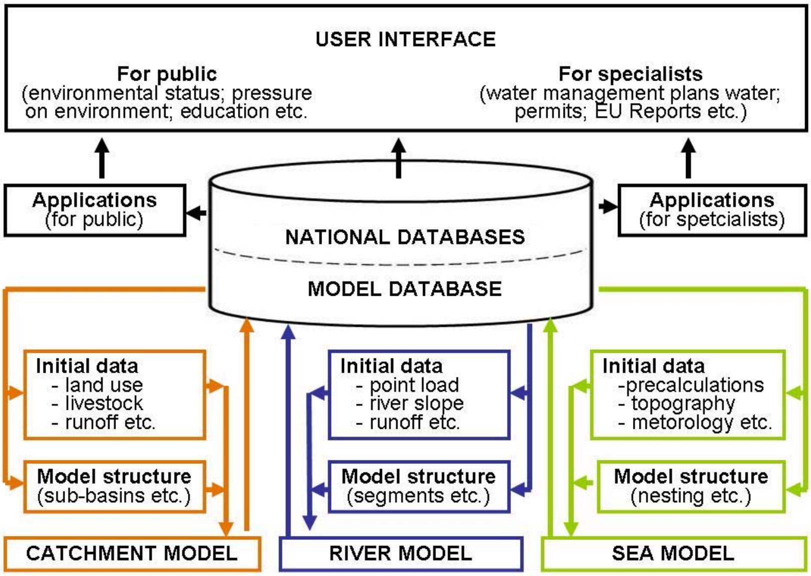


Figure 3. Model dependencies

When considering what sort of data to use, it cannot be decided upon simply, as each different type of data has a specific price tag – the cost of time and money. The difficult ordeals are reaching conclusions to which scale of data is required at the moment (when a query is made from our information system) and which data to collect and use in the future to conduct further analysis. For example, the Estonian Marine Systems Institute, who work with oceanographic models, have around 30 measurement stations in Estonian marine waters, which cost many thousands of euros per month, only to maintain for collecting data, not including storage and processing. Necessary data for models can be generated on-demand, by extending data rows (1) for the particular model and its information needs, such as estimating river flows:

$$\left.\begin{aligned} & Q(k,b) = \sum_{i=1}^{n}(A_i - kB_i - b)^2 \\ & \hat{k} = \frac{\sum_{i=1}^{n} A_i B_i - \frac{1}{n}\sum_{i=1}^{n} Bi \sum_{j=1}^{n} A_j}{\sum_{i=1}^{n}(B_i^2) - \frac{1}{n}(\sum_{i=1}^{n} B_i)^2} \\ & \hat{b} = A_i - \hat{k}B_i \end{aligned}\right\} \qquad (1)$$

A – measurement on river A; B measurement on river B; Q, k, b – river flows.

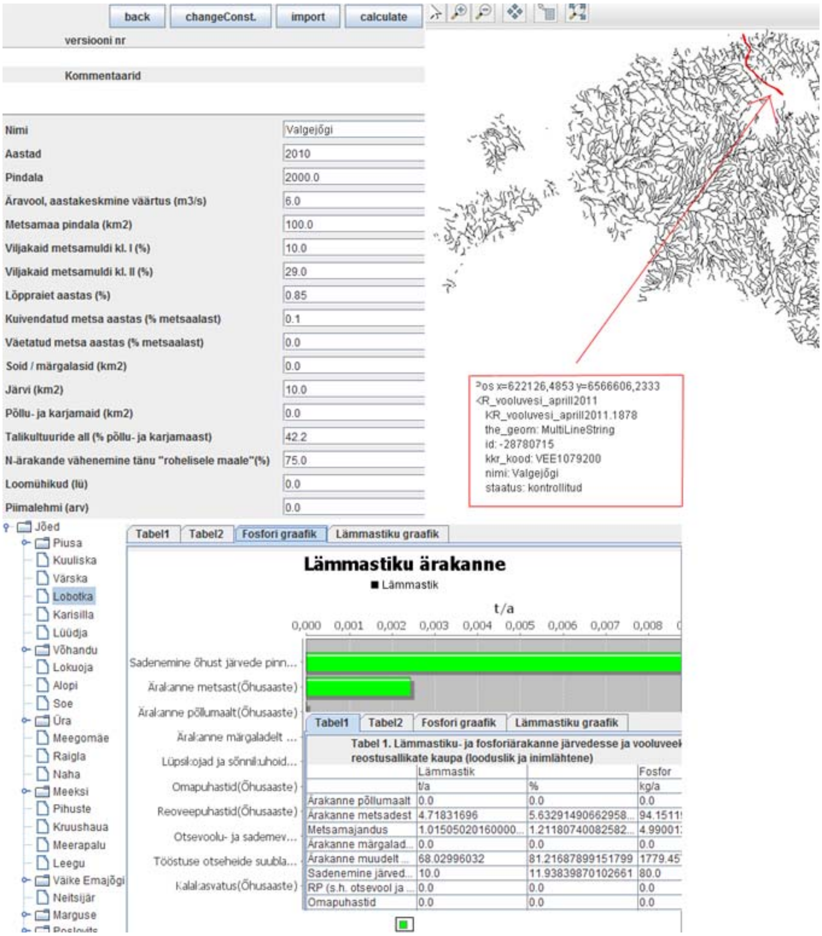
3. Environmental Data Applications

Monitoring data often needs to be processed due to various reasons and the EWIS means to provide processing capabilities. Since a large amount of information is needed for generating water management plans, models are important. First, the observation network density is never sufficient to assess the status of all water bodies – our databases include more than 6600 surface water bodies. Second, observations can't describe of the environmental consequences of planned activities. Third, complex processes in the environment give only integrated outcome of all affecting factors. It is clear that the selected model must be suitable for the set challenges. E.g., for catchment area modeling we selected a very simple model, developed by Tord Wennerblom for the county of Älvsborg in Sweden [6]. This model has been converted from Excel form into a fully interactive, programmatically correct Java application, and can be launched either as an applet or a desktop program. An example simulation for a hydrological specialist end-user is shown in Figure 4. Sample code from the inner workings of the GeoTools framework we are using to build the EWIS interactive map is as follows:

```
/**
*The last line is responsible for excecuting the query, the query result will be stored in featureCollection,
*which can later be used to edit or process geographical feature data.
*/
private void queryFeatures() throws Exception {
    String featureTypeName = (String) featureTypeCBox.getSelectedItem();
    FeatureSource source = dataStore.getFeatureSource(typeName);
    FeatureType schema = source.getSchema();
    String featureName = schema.getGeometryDescriptor().getLocalName();
    Filter filter = CQL.toFilter(text.getText());
    DefaultQuery query = new DefaultQuery(schema.getName().getLocalPart(), filter,
        new String[] { featureName });
    SimpleFeatureCollection featureCollection = source.getFeatures(query);
}
```

The preceding code executes a query to a database of our selection, returning parameters the model needs, based on map selection. Another snippet shows how this would work using SQL, when querying straight from tables, if the map query has exhausted itself and additional queries based on geographical features from the map must be made to gather further information, e.g.:

```
private void queryFeaturesFromTable() throws Exception {
    ResultSet rs = st.executeQuery("SELECT DISTINCT v.id as id, v.nimi as nimi " +  "FROM
public_sr_programm as s_p, public_obj_programm as o_p, public_seirejaam as sj, public_veekogu as v " +
"WHERE s_p.id=o_p.programm_id AND o_p.obj_id=sj.id AND v.id=sj.veekogu_id " + "AND s_p.id = " +
temp.id + " ORDER BY v.nimi");
}
```



versiooni nr

Kommentaarid

Nimi	Valgejõgi
Aastad	2010
Pindala	2000.0
Äravool, aastakeskmine väärtus (m3/s)	6.0
Metsamaa pindala (km2)	100.0
Viljakaid metsamuldi kl. I (%)	10.0
Viljakaid metsamuldi kl. II (%)	29.0
Lõppraiet aastas (%)	0.85
Kuivendatud metsa aastas (% metsaalast)	0.1
Väetatud metsa aastas (% metsaalast)	0.0
Soid / märgalasid (km2)	0.0
Järvi (km2)	10.0
Põllu- ja karjamaid (km2)	0.0
Talikultuuride all (% põllu- ja karjamaast)	42.2
N-ärakande vähenemine tänu "rohelisele maale"(%)	75.0
Loomühikud (lü)	0.0
Piimalehmi (arv)	0.0

Pos x=622126,4853 y=6566606,2333
KR_vooluvesi_aprill2011
KR_vooluvesi_aprill2011.1878
the_geom: MultiLineString
id: -28780715
kkr_kood: VEE1079200
nimi: Valgejõgi
staatus: kontrollitud

Tabel 1. Lämmastiku- ja fosforiärakanne järvedesse ja vooluveek
reostusallikate kaupa (looduslik ja inimlähtene)

	Lämmastik		Fosfor
	t/a	%	kg/a
Ärakanne põllumaalt	0.0	0.0	0.0
Ärakanne metsadest	4.71831696	5.63291490662958...	94.1511
Metsamajandus	1.01505020160000...	1.21180740082582...	4.99001
Ärakanne märgalad...	0.0	0.0	0.0
Ärakanne muudelt ...	68.02996032	81.21687899151799	1779.45
Sadenemine järved...	10.0	11.93839870102661	80.0
RP (s.h. otsevool ja ...	0.0	0.0	0.0
Omapuhastid	0.0	0.0	0.0

Figure 4. Wennerblom model implemented in Java

The calculations of the Wennerblom model are based on five forest-dominated rivers in Northern Svealand and Norrland (central and northern Sweden). The runoff from the ground in forests has been calculated as functions of discharges:

$$Q = f\,(\text{NH4-N, NO3-N, Org-N, Tot-N, Tot-P}) \qquad (2)$$

The units are, for each compound, kg/ha per year, whereas Q (mm/year) is the runoff (2). Due to the relative similarity between Estonian and Swedish climate and landscape, the Wennerblom model can easily be applied to Estonian surface waters using calibrated coefficients by the EEIC in the future. The Wennerblom model was chosen due to its simplicity, as it is easier to implement a simpler model to test the unification of our databases. Since it is a simpler model, its credibility might be lower than a complex model's, but work is underway in comparing results from the Wennerblom model with the Soil and Water Assessment Tool (SWAT) results, to see which offers more accurate modeled data.

4. Conclusion

By utilizing the data provided by the EEIC, we wish to develop a system, EWIS, to provide the general public with information about the environmental situation of Estonia's surface waters. Since the EEIC stores its data in many heterogeneous sources, data integration is a serious issue. To combat the discrepancies in data, we have developed a mediated data management system, nicknamed ROOMA, which is written in Java using the Hibernate framework. The Open Data used to answer an end-user's [7] query, even when processed, will remain as Open Data and will not be monetized in any way. To accomplish our goals, we use an amalgam of models, which extend data rows on-demand, also reducing required intermediate data storage for calculations.

It is our aim to deal with the following issues in the nearest time:

- Move to cloud-based data storage and service hosting.
- Expand the possibilities of the EWIS data integration by adding coastal models and additional databases.
- Integration into the Estonian national data exchange grid, X-road.

References

[1] *Aarhus Convention*. *Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters*. UNECE, 1998 [cited 2012 February 1]. Available from: http://www.unece.org/env/pp/treatytext.htm.

[2] U. Vallner, *Cooperation Capability of Estonian Open Data*. RIA, 2011 [cited 2012 February 2]. Available from: https://www.ria.ee/public/Programm/Tark_e_riik_2011/Avaandmete_teabepaev_31.01.12/1_Riigi_avaandmete_koosvoime_Uuno_Vallner_2012-01-31.pdf.

[3] A. Kalja, *The Data Exchange Layer X-Road in 2008*. Ministry of Economic Affairs and Communications, 2009 [cited 2012 February 27]. Available from: http://www.riso.ee/en/files/Yearbook2008/pdf/yearbook2008.pdf.

[4] T. Nurmela, *Cloud Computing - International Standardization and Industry Consortium*. Cybernetica AS, 2011 [cited 2012 February 10]. Available from: http://www.cyber.ee/publikatsioonid/30-info-ja-teabepaevade-ettekanded/cloud_computing_Nurmela.pdf.

[5] V. Viies, P. Ennet, J. Aigro, H. Kinks, R. Kullamaa, O. M. Ozolit, and A. Salula, *Database Interface Compatibility in the Estonian Water Information System*. Cybernetica AS, 2011 [cited 2012 February 1]. Available from: http://www.cyber.ee/publikatsioonid/30-info-ja-teabepaevade-ettekanded /Eesti%20Vee...pdf.

[6] H. Lindström, J. Gunnarson, T. Wennerblom, and H. Kvarnäs, Implementing sustainable water regimes. In: L.C. Lundin (ed.), *Sustainable Water Management in the Baltic Sea Basin. Book III. River Basin Management*, Ditt Tryckeri i Uppsala AB, 2000, 221–229.

[7] V. Viies, P. Ennet, J. Aigro, H. Kinks, R. Kullamaa, O. M. Ozolit, and A. Salula, Water Information System for Estonia. In: *Material of 7th ATINER Inernational Conference on Computer Science & Information Systems*, Greece, Athens, 2011.

Hypermodelling Live
OLAP for Code Clone Recommendation

Tim FREY and Veit KÖPPEN
Otto-von-Guericke-University Magdeburg, Germany

Abstract. Code bases contain often millions lines of code. Code recommendation systems ease programming by proposing developers mined and extracted use cases of a code base. Currently, recommender systems are based on hardcoded sets what makes it complicate to adapt them. Another research area is adaptable live detection of code clones. We advance clone detection and code recommender systems by presenting utilization of our Hypermodelling approach to realize an alternative technique. This approach uses Data Warehousing technology that scales for big data and allows for flexible and adaptable queries of source code. We present the generic idea to advance recommendation and clone detection based on queries and to evaluate our application with industry source code. Consequently, recommender systems and clone detection can be customized with flexible queries via Hypermodelling. This enables further research about more complex clone detection and context sensitive code recommendation.

Keywords. Data warehousing, software engineering, hypermodelling

Introduction

Code recommender systems advance integrated development environments. They are based on the idea to extract and mine information from code bases to generate recommendations. Recommendation data contains information, which method calls occur commonly together or which methods of a super class get overwritten [1]. This data is compared to current coding of a developer and proposals are offered. The extraction and mining process limits recommendation to be easily adjusted to specific requirements. For instance, it is desirable to have recommendation information available for diverse APIs and also for an own project. Furthermore, project requirements often differ. In one project, it is required that the recommendation code base just comes out of a specific project and in another setting all available code should be used for recommendation. However, this type of flexible recommendation is currently not available. One main reason for this may be the immense size of modern code bases, resulting in difficulties to adjust the extraction process: It is necessary to scan the code base for different recommendation information every time.

Another challenge is the detection of code clones [2, 3]. Thereby, code bases are scanned for duplicates. One main challenge is the different type of code bases and clone detection methods. Sometimes, a certain package should be excluded because replicas are allowed. Furthermore, clones may be exact duplicates of a code fragment or similar pieces of code. Hence, clone detection faces the challenge to provide an easy adjustable infrastructure that allows detecting different kinds of clones and different code base configurations.[1]

Altogether, recommender systems and clone detection face the challenge to provide an adaptable infrastructure for large code bases that allow by flexible means to detect code clones and recommendations. In order to overcome these current limitations, we propose to use our Hypermodelling approach [4, 5, 6] for flexible code recommendation and clone detection. Through this approach both techniques can be covered with Data Warehouse (DW) and Online Analytical processing (OLAP) technology [7] that scales well for large data sets and is easily adjustable for multi-dimensional queries. Therefore, our contribution is to describe, how the Hypermodelling approach can be used to advance recommender systems and clone detection at the same time. We also provide an evaluation, in which we demonstrate the approach on a real world source code excerpt.

The remainder of the paper is structured as follows: First, we recap important facts about our Hypermodelling approach. Next, we describe the general approach, how Hypermodelling can be used to detect clones and recommend code fragments. Afterwards, we describe the approach with a concrete example. In Section 4, we refer to related work and point out differences.

1. Hypermodelling

Hypermodelling is the idea to combine program analysis and DW. A more detailed description of the Hypermodelling approach is available in [4].[2] In [5, 6] different reporting possibilities of this approach are presented.

DW systems are an integrative component in business computing [7].[3] They are used to assemble data of different sources together. The integrated data are arranged into multi-dimensional data structures, i.e., data cubes, which serve as base for queries [4]. Queries can be used to aggregate different measures and dimensions (within their hierarchies) that occur in the data. For instance, sales for an employee can be computed for a given time period. Thereby, this query aggregates the region, the products (sales) and the time in relation to financial indicators. Likewise, hierarchies can be abstracted. For instance, the region can be split into continents, countries, counties, as well as cities and the aggregates are associated with the distinct sales for those. This can be done for other hierarchies, e.g., customer group, year, or department. Generally, the idea is that different aggregations enable detailed investigations. With Hypermodelling, we introduce the idea that programming elements, like annotations or classes, are similar to data that are used within a DW. For instance, classes are defined within a package hierarchy. Annotations are associated with classes and their members. They are also defined in their own package. This is like the association of a salesman to a region, time period, and revenues. Hierarchies in code are similar to hierarchies of region or time. All together, we load source code into a DW and realize associations of classes, their inheritance, packages, and annotations as a DW cube and execute queries on it. For this paper, we combine cubes that are used in [4, 5, 6]. We present the results of exemplarily queries on our aggregated cube as reports in this paper, see Section 3.

[1] Note, clone detection methods can be configured and live detection is possible [3]. However, it still is a challenge. Thus, we present alternative approach as complementary subsidiary to code recommendation.

[2] http://hypermodelling.com

[3] Note, there are also open source Data Warehouse solutions available. For instance: http://pentaho.com.

2. Clone Recommendation with Hypermodelling

First, we describe the overall idea to use OLAP queries for code analysis. Afterwards, we refer to recommendation and clone detection.

2.1. Overall Idea

We propose a query based approach to advance recommendation and clone detection. We use OLAP queries and our data warehouse approach, out of the size of modern code bases. Additionally, DW technology has many best practices and tuning methods at hand. Queries allow flexibility, adaptability, and DW technology scales for big data. Thus, our live query approach should help to avoid clones and give recommendations at the same time.

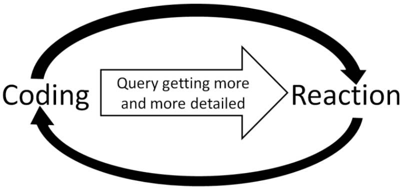


Figure 1. Query refinement throughout the coding process

A developer encodes functionality (coding) and in the background, her written code is used to execute an OLAP query against the Code-DW that contains coded structure and elements. For instance, such code can be parameters of methods. Then, the query result gets presented to the developer. This may be in the form of recommendation or in a notification of a code clone. Hence, this query result influences ongoing coding process. From the result, the programmer goes on and encodes more functionality. With this additional functionality, again a query can be executed that contains more information than the first. Every time, as more and more code exists, the query is more and more detailed. If the programmer finishes her current coding, the query is quite detailed with all code fragments that are belonging to a method. If another method shares enough similarities, the corresponding code is presented to the developer and she has to decide if a code clone exists.

Currently, recommendation is not done live and neither is based on a query approach. Even more, code recommenders are inflexible and do not support slicing and dicing data for specific needs. This is often likewise the same for clone detection. So, with our Hypermodelling approach to load code into a DW, every query could be customized to meet specific requirements or just be sliced by a specific viewpoint.

2.2. Recommendation

Code recommendation systems mine facts that plenty of programmers have done out of a code base. Imagine, a developer overrides a method. The written method is compared with the recommendation data and she gets proposed which methods were called by others that did an override of this method. Exactly the same information can be revealed with an OLAP query. This information can be presented to a developer to recommend her what others did.

What we describe is a typical application of a code recommender system. Therefore, we propose a query process for recommender systems in Figure 1. A developer encodes functionality and queries containing this encoded functionality are done. These queries reveal similar code artifacts and most things that others have done in those are proposed to a developer. The developer continues to encode functionality. This encoded functionality can be added as refinement to the query to get more detailed information on what most others did in a similar situation. With these queries, the recommendation data can be generated live and adapted easily. The Hypermodelling approach ensures that different code bases can be loaded into a DW. This allows us to reuse infrastructure that is designed for big data.

2.3. Clone Detection

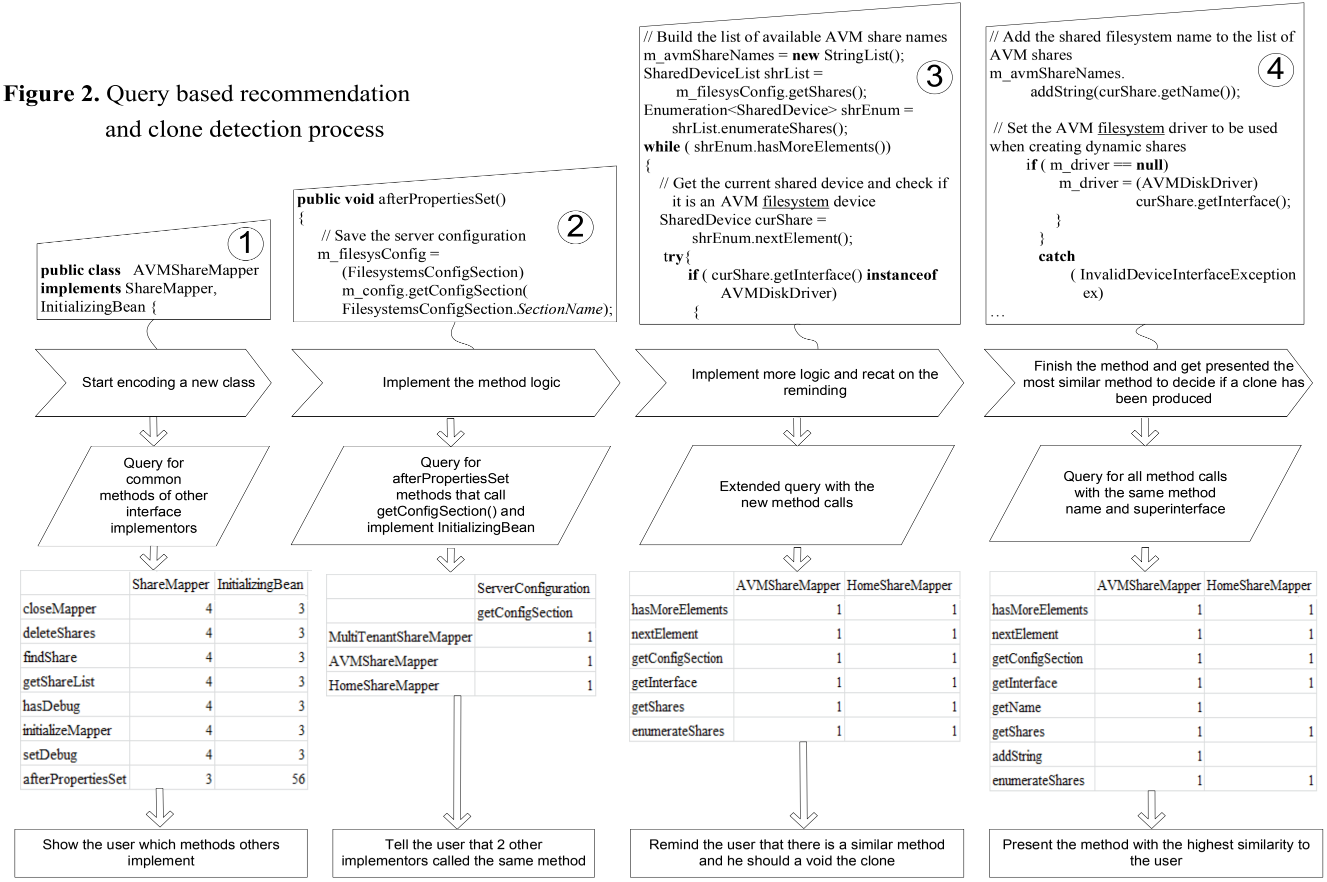
Similarly to recommendation, we also realize live clone detection. Imagine a developer encodes a method, sharing a high similarity or equality to another method. Today, she continues her work. However, in this moment, she has knowledge at hand what she has programmed and can easily compare it with potential clones. This step fits perfectly into workflow and potential clones can be avoided easily. We propose to use our Hypermodelling approach to detect code clones live though a query based approach as described in Section 2.2. Through queries the detection process can be easily adjusted to project specific requirements.

For clone detection, imagine a developer is encoding functionality in a method. She creates the method declaration with all method parameters and then she encodes logic into the method. Methods of objects are called and other constructs are realized. Like for recommenders, regularly queries are executed to determine duplicates. If similarity is too high, code is presented to the developer so that she can decide if she has produced a clone. We show this process in Figure 1, that also describes a clone detection cycle.

3. Evaluation

In order to evaluate the application ability of recommendation and clone detection based on queries, we select a method of a program that we describe in [6]. Data in our DW and queries on a real project demonstrate that our approach enables recommendation not only with prepared data. We depict a class (*AVMShareMapper*) that implements two interfaces that are implemented by other classes of the application. Thereby, we query if others also implement these interfaces to ensure a valid example that shares similarities with other classes. We select the *afterpropertiesSet* method for investigation and divide the method in four different parts. For every part, we execute queries to simulate how a developer would encode this method and queries would be executed in the background. Our scenario is mainly based on live clone detection. However, the same approach of queries can be used for recommendation.

Figure 2 shows the extended coding process that can be supported by queries. We imagine the coding process there as follows: In the first step a developer encodes the class body. Then, she goes on and encodes step by step a method. We split the *afterpropertiesSet* method and arranged it above, corresponding to the process of a developer. We describe exemplary queries in natural language beneath the process. Those queries can be executed with the DW query languages from the development environment in the background, while the developer is coding. Behind queries is their



```
public class   AVMShareMapper
implements ShareMapper,
InitializingBean {
```

1
Start encoding a new class
Query for common methods of other interface implementors

	ShareMapper	InitializingBean
closeMapper	4	3
deleteShares	4	3
findShare	4	3
getShareList	4	3
hasDebug	4	3
initializeMapper	4	3
setDebug	4	3
afterPropertiesSet	3	56

Show the user which methods others implement

```
public void afterPropertiesSet()
{
    // Save the server configuration
    m_filesysConfig =
        (FilesystemsConfigSection)
        m_config.getConfigSection(
        FilesystemsConfigSection.SectionName);
```

2
Implement the method logic
Query for afterPropertiesSet methods that call getConfigSection() and implement InitializingBean

	ServerConfiguration
	getConfigSection
MultiTenantShareMapper	1
AVMShareMapper	1
HomeShareMapper	1

Tell the user that 2 other implementors called the same method

```
// Build the list of available AVM share names
m_avmShareNames = new StringList();
SharedDeviceList shrList =
      m_filesysConfig.getShares();
Enumeration<SharedDevice> shrEnum =
     shrList.enumerateShares();
while ( shrEnum.hasMoreElements())
{
   // Get the current shared device and check if
     it is an AVM filesystem device
   SharedDevice curShare =
        shrEnum.nextElement();
   try{
       if ( curShare.getInterface() instanceof
             AVMDiskDriver)
       {
```

3
Implement more logic and recat on the reminding
Extended query with the new method calls

	AVMShareMapper	HomeShareMapper
hasMoreElements	1	1
nextElement	1	1
getConfigSection	1	1
getInterface	1	1
getShares	1	1
enumerateShares	1	1

Remind the user that there is a similar method and he should a void the clone

```
// Add the shared filesystem name to the list of
AVM shares
m_avmShareNames.
       addString(curShare.getName());

 // Set the AVM filesystem driver to be used
when creating dynamic shares
      if ( m_driver == null)
           m_driver = (AVMDiskDriver)
                        curShare.getInterface();
         }
      }
      catch
           ( InvalidDeviceInterfaceException
             ex)
…
```

4
Finish the method and get presented the most similar method to decide if a clone has been produced
Query for all method calls with the same method name and superinterface

	AVMShareMapper	HomeShareMapper
hasMoreElements	1	1
nextElement	1	1
getConfigSection	1	1
getInterface	1	1
getName	1	
getShares	1	1
addString	1	
enumerateShares	1	1

Figure 2. Query based recommendation and clone detection process

result or at least an excerpt. At the bottom, possible ensuing actions are described. In the following, we go through the process of Figure 2:

First (1), a developer starts encoding the class and implements the *ShareMapper* and the *InitializingBean* interface. These two interfaces are used to create a query for the most common method names of the classes that implement one of the interfaces.[4] The result shows the amount of children types with the same method name. The result is sorted following the occurrence of method names of *ShareMapper* implementers. It would be possible to sort the result after the occurrence in *InitializingBean* or to merge most common method names in both interface implementers. Anyway, the result information can be used to show developers which methods other developers implemented by extending a certain interface. For our scenario, we imagine that developers see plenty of times the *afterPropertiesSet* method is implemented and start encoding this. Thereby, it can be recognized that a developer uses a method name based on the *InitializingBean* and the following queries can be specialized on this interface. To give a better impression about the technique, we present the query in Listing 1. The parent class or interface is named *ParentType* and the CodeStructure is the OLAP cube. The query is based on multi-dimensional expressions standard[5] and shows that the amount of methods for children is computed for extenders of the *ShareMapper* and *InitializingBean*.

```
SELECT { [ParentType].[Name].&[ShareMapper], [Parent].[Name].&[InitializingBean] } ON COLUMNS ,
{ [Method].[Name].[All].CHILDREN } ON ROWS
FROM [Code-Structure]  WHERE ( [Measures].[Method-Count] )
```

Listing 1. Query for method names of interfaces children

In the second step (2), the developer starts encoding logic of the *afterPropertiesSet* method. She calls a method of what is used to refine the former query. It is enriched with information which methods are called to reveal which types also obtain same methods, implement the interface and have the same method name. Such kind of similarity can be an indicator that the developer produces a clone. Furthermore, maybe the developer is putting effort and thoughts into implementing a method that is already implemented. Therefore, a developer gets presented other methods that share a high similarity. If a user rejects the proposals and wants to encode further functionality, the information about similar methods can also be used to present methods that are called in the similars. We show the corresponding query in Listing 2. The result types of this query that share a similarity (*MultiTenantShareMapper*, *HomeShareMapper*) can also be used in another query to generate recommendation information. Exemplarily, such a recommender query is shown in Listing 3. The called methods of two similar types (*MultiTenantShareMapper*, *HomeShareMapper*) are computed based on a query. This can be used to propose a user which other methods are called by other developers in a similar situation.

```
SELECT { [CalledMethod].[Name].&[getConfigSection]} ON COLUMNS ,
{[Type].[Name].[All].CHILDREN} ON ROWS   FROM [Code-Structure]
WHERE ( [Measures].[Method Calls],[ParentType].[Name].&[InitializingBean],
[Method].[Name].&[afterPropertiesSet] )
```

Listing 2. Determining similar code

```
SELECT NON EMPTY { [Measures].[Method Calls  Anzahl] } ON COLUMNS ,
{ [Called Method].[Name].[All].CHILDREN } ON ROWS
FROM [Code-Structure]
WHERE ( [Type].[Name].&[HomeShareMapper], [Parent - Type].[Name].&[InitializingBean],
[Method].[Name].&[afterPropertiesSet] )
```

Listing 3. Determining what similar code did for recommendation

[4] Note, a query can be sliced by a project or a specific package. Plenty of customizations are thinkable.

[5] We show a few queries to give an impression of the query language. For further information, see the MDX language reference: http://msdn.microsoft.com/en-us/library/ms145595.aspx / http://xmlforanalysis.com

Succeeding (3), a developer encodes more functionality. In the background the query is extended to compare method calls of the current method with the ones of other extenders. It is like a fail save detector that tries to uncover if there is a clone produced. With every new method call, the query is extended and executed again. After six shared method calls and the super interface, enough indicators are collected to remind the user with a traffic light or a pop up about similarity.[6]

Finally, in the forth step (4), the user finishes the method. This is recognized in the background and a complete query for all method calls is executed. Out of the high similarity of the method calls, it is assumed that the method logic is maybe "externalizable" and customizable through parameters. However, the easiest way is to present methods with a high similarity to the user and let him decide if this is a clone.

4. Related Work

Our prior works describes a general approach how code structure can be processed with a DW and queries [4, 5, 6]. In this paper, we focused on a concrete application.

Others describe different clone detection methods [2]. One clone detection method [3] proposes live clone detection with a client-server architecture. In comparison to our approach, the described detection [3] is not customizable. Another one, CloneDetective [8], offers an advanced framework and tool chain for clone detection, which is especially geared towards flexibility of clone detection research. Hypermodelling, on the contrary, targets to utilize DW technology. In general, we see the use of DWs as an addition to known clone detection mechanisms and not as competitor. DWs often already exist in enterprises where they are used for business applications. Therefore, our approach makes a reuse of DW technology for clone detection possible. Hence, further research should determine in detail if and which of those other clone detection approaches can be realized with DW technology.

Other related work can be found in the area of code recommender systems [1]. They provide methods to mine data out of code bases and generate recommendations out of it. The whole recommendation is hard wired and fixed. Further, no scaling technology for live recommendations with a client server infrastructure is proposed. With our approach, through DW query based recommendations a more flexible and adjustable infrastructure is at hand. Therefore, we see the emerging need to investigate DW technology further to advance the recommender systems.

5. Conclusion and Future Work

We described the problem of large code bases and the inflexibility of current code recommendation and challenges of clone detection systems. We proposed to overcome these limitations through a query based approach that uses DW technology. Our approach is evaluated and its application ability is shown by queries to a real code base. So, the next generation recommendation and clone detection techniques can be based on DW technology. Related work reveals possible synergies with other research and indicates that DW technology is a promising area to advance software engineering.

[6] Note, at this moment such kind of queries can also be used to propose method calls like in step 2.

In general, we see the need to describe our method in more details. Additionally, the use of DW technology enables further research on DW based clone detection. We showed the capability to find "full method clones". However, clones can also be copied code fragments within different methods. Therefore, further investigations can focus on identification of queries for these clones. Thereby, we see ways to compute similarity indicators based on queries and the clone granularity level (method or fragment based) as important questions. The precision of our technique is fixed to method calls, discarding method parameters. Hence, the same called methods that take different parameters and are considered as clones. For that reason, the possibility to enhance or adjust the precision of our queries through additional facts should be considered. Therefore, we see the need to work together with industry developers to evaluate which level of precision and granularity is desired in practice.

Furthermore, we connect code recommendation with clone detection. Therefore, our work makes it possible to regard both areas together and investigate possible synergies in the future. This connection is an additional difference to previous research. Generally, our approach shows a rudimentary and first scenario with primitive queries. More complex queries, scenarios, and areas are of interest for further investigations. For instance, the area of refactoring is also near clone detection and relations to it can be investigated. We also see an advanced trail by integrating the context (e.g., the package, the prior studied code, or the task) wherein a developer encodes functionality to advance recommendation systems. Currently, code recommenders are dull and based on the same rule set. Our dynamic query based approach enables further research, how recommendations need to be altered and adjusted to different contexts of a developer to respect his current programming tasks within the recommendation.

References

[1] M. Bruch, M. Mezini, and M. Monperrus, Mining subclassing directives to improve framework reuse, In: *Proceedings of the 7th IEEE Working Conference on Mining Software Repositories*, IEEE, 2010, 141-150.

[2] C. K. Roy, J. R. Cordy, and R. Koschke, Comparison and evaluation of code clone detection techniques and tools: a qualitative approach, *Science of Computer Programming* **74**(7) (2009), 470-495.

[3] T. Yamashina, H. Uwano, K. Fushida, Y. Kamei, M. Nagura, S. Kawaguchi, and H. Iada, *Shinobi: a Real-Time Code Clone Detection Tool for Software Maintenance*, Technical Report NAIST-IS-TR2007011. Graduate School of Information Science, Nara Institute of Science and Technology, 2008.

[4] T. Frey, V. Köppen, and G. Saake, Hypermodelling – introducing multi-dimensional concern reverse engineering**.** In: *2nd International ACM/GI Workshop on Digital Engineering (IWDE),* Germany, 2011, 58-66.

[5] T. Frey, Hypermodelling for drag and drop concern queries. In: *Proceedings of Software Engineering 2012 (SE2012)*, Gesellschaft für Informatik (GI), Berlin, Germany, 2012, 107-118.

[6] T. Frey and V. Köppen, Exploring software variance with hypermodelling – an exemplary approach. In: S. Jähnichen, A. Küpper, S. Albayrak, editors, *Software Engineering 2012: Fachtagung des GI-Fachbereichs Softwaretechnik,* Berlin, Germany, 2012, 121-140.

[7] W. H. Inmon, *Building the Data Warehouse*. 4th ed. J.Wiley & Sons, New York, USA, 2005.

[8] E. Juergens, F. Deissenboeck, and B. Hummel, CloneDetective – a workbench for clone detection research. In: *Proceedings of the 30th International Conference on Software Engineering*, IEEE, 2009, 603-606.

Efficient Subset and Superset Queries

Iztok SAVNIK
Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, 5000 Koper, Slovenia

Abstract. The paper presents index structure for storing and querying sets called $SetTrie$. Besides the operations $insert$ and $search$ defined for ordinary tries, we introduce the operations for retrieving subsets and supersets of a given set from $SetTrie$ tree. The performance of operations is analysed empirically in a series of experiments. The analysis shows that sets can be accessed in $\mathcal{O}(c * |set|)$ time where $|set|$ represents the size of parameter set. The constant c is up to 5 for subset case and approximately 150 in average case for the superset case.

Keywords. Containment queries, indexes, access methods, databases

Introduction

Set containment queries are common in applications based on object-oriented or object-relational database systems. Relational tables or objects from collections can have *set-valued attributes* i.e. the attributes that range over sets. Set containment queries can express either selection or join operation based on set containment condition [5,10,3].

In this paper we propose an index structure $SetTrie$ that implements efficiently the basic two types of set containment queries: subset and superset queries. We give the presentation of the proposed data structure, the operations defined on $SetTrie$ and thorough empirical analysis.

Let us first give a description of subset and superset operations in more detail. Let U be a set of ordered symbols. The subsets of U are denoted as *words*. Given a set of words S and a subset of U named X, we are interested in the following queries.

1. Is X a subset of any element from S?
2. Is X a superset of any element from S?
3. Enumerate all Y in S such that X is a subset of Y.
4. Enumerate all Y in S such that X is a superset of Y.

$SetTrie$ is a tree data structure similar to $trie$ [6]. The possibility to extend the performance of usual $trie$ from membership operation to subset and superset operations comes from the fact that we are storing *sets* and not the *sequences* of symbols as for ordinary tries. In the case of sets the ordering of symbols in a set is not important as it is in the case of text. As it will be presented in the paper the ordering of set elements can be exploited for the efficient implementation of containment operations.

We analyse subset and superset operations in two types of experiments. Firstly, we examine the execution of the operations on real-world data where sets represents words from the English dictionary. Secondly, we have tested the operations on artificially gen-

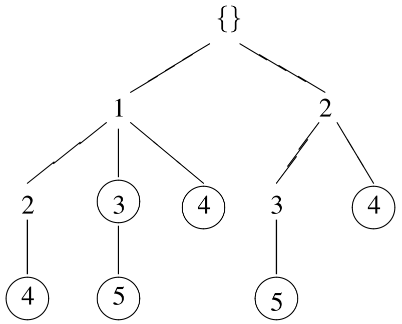


Figure 1. Example of $SetTrie$

erated data. In these experiments we tried to see how three main parameters: the size of words, the size of $SetTrie$ tree and the size of test-set, affect the behavior of the operations.

The paper is organized as follows. The following section presents the data structure $SetTrie$ together with the operations for searching the subsets and supersets in a tree. The Section 2 describes the empirical study of $SetTrie$. We present a series of experiments that measure the behavior of operations and the size of data structure. The related work is presented in Section 3. We give the presentation of existent work on set-valued attributes and containment queries as well as related work on trie and Patricia tree data structures. Finally, the overview and conclusions are given in Section 4.

1. Data Structure $SetTrie$

$SetTrie$ is a tree composed of nodes labeled with indices from 1 to N where N is the size of the alphabet. The root node is labeled with $\{\}$ and its children can be the nodes labeled from 1 to N. A root node alone represents an empty set. A node labeled i can have children labeled with numbers greater than i. Each node can have a flag denoting the last element in the set. Therefore, a set is represented by a path from the root node to a node with flag set to true.

Let us give an example of $SetTrie$. Figure 1 presents a $SetTrie$ containing the sets $\{1, 3\}$, $\{1, 3, 5\}$, $\{1, 4\}$, $\{1, 2, 4\}$, $\{2, 4\}$, $\{2, 3, 5\}$. Note that flaged nodes are represented with circles.

Since we are dealing with sets for which the ordering of the elements is not important, we can define a syntactical order of symbols by assigning each symbol a unique index. Words are ordered by sequences of indices. The ordering of words is exploited for the *representation* of sets of words as well as in the *implementation* of the above stated operations.

$SetTrie$ is a tree storing a set of words which are represented by a path from the root of $SetTrie$ to a node corresponding to the indices of elements from words. As with tries, prefixes that overlap are represented by a common path from the root to an internal vertex of $SetTrie$ tree.

The operations for searching subsets and supersets of a set X in S use the ordering of U. The algorithms do not need to consider the tree branches for which we know they do not lead to results on the basis of the ordering of word symbols. The search space for a given X and tree representing S can be seen as a subtree determined primarily by the search word X but also with the search tree corresponding to S.

1.1. Operations

Let us first present a data structure for storing *words*, that is, the sets of symbols. Words are stored in a data structure $Word$ representing ordered sets of integer numbers.

The users of $Word$ can scan sets using the following mechanism. The operation $word.gotoFirstElement$ sets the current element of word to the first element of ordered set. Then, the operation $word.existsCurrentElement$ checks if word has the current element set. The operation $word.currentElement$ returns the current element, and the operation $word.gotoNextElement$ goes to the next element in the set.

Let us now describe the operations of the data structure $SetTrie$. The first operation is insertion. The operation $insert$(root,word) enters a new $word$ into the $SetTrie$ referenced by the root $node$. The operation is presented by Algorithm 1.

Algorithm 1 insert($node$, $word$)

```
if ($word$.existsCurrentElement) then
    if (exists child of $node$ labeled $word$.currentElement) then
        $nextNode$ = child of $node$ labeled $word$.currentElement;
    else
        $nextNode$ = create child of $node$ labeled $word$.currentElement;
    end if
    insert($nextNode$, $word$.gotoNextElement)
else
    $node$'s flag_last = true;
end if
```

Each invocation of operation $insert$ either traverses through the existing tree nodes or creates new nodes to construct a path from the root to the flagged node corresponding to the last element of the ordered set.

The following operation $search$(node,word) searches for a given $word$ in the tree $node$. It returns true when it finds all symbols from the word, and false as soon one symbol is not found. The algorithm is shown in Algorithm 2. It traverses the tree $node$ by using the elements of ordered set $word$ to select the children.

Let us give a few comments to present the algorithm in more detail. The operation have to be invoked with the call $search$(root,set.gotoFirstElement) so that $root$ is the root of the $SetTrie$ tree and the current element of the $word$ is the first element of $word$. Each activation of $search$ tries to match the current element of $word$ with the child of $node$. If the match is not successful it returns $false$ otherwise it proceeds with the following element of $word$.

The operation $existsSubset$(node,word) checks if there exists a subset of $word$ in the given tree referenced by $node$. The subset that we search in the tree has fewer elements than $word$. Therefore, besides that we search for the exact match we can also skip one or more elements in $word$ and find a subset that matches the rest of the elements of $word$. The operation is presented in Algorithm 3.

Algorithm 2 search($node$, $word$)

```
if ($word$.existsCurrentElement) then
   if (there exists child of $node$ labeled $word$.currentElement) then
      matchNode = child vertex of $node$ labeled $word$.currentElement;
      search($matchNode$, $word$.gotoNextElement);
   else
      return false;
   end if
else
   return ($node$'s last_flag $==$ true) ;
end if
```

Algorithm 3 existsSubset(node,set)

```
 1: if ($node$.last_flag $==$ true) then
 2:    return true;
 3: end if
 4: if (not $word$.existsCurrentElement) then
 5:    return false;
 6: end if
 7: found = false;
 8: if ($node$ has child labeled $word$.currentElement) then
 9:    $nextNode$ = child of $node$ labeled $word$.currentElement;
10:    $found$ = existsSubset($nextNode$, $word$.gotoNextElement);
11: end if
12: if (!found) then
13:    return existsSubset($node$,$word$.gotoNextElement);
14: else
15:    return true;
16: end if
```

Algorithm 3 tries to match elements of $word$ by descending simultaneously in tree and in $word$. The first IF statement (line 1) checks if a subset of $word$ is found in the tree i.e. the current node of a tree is the last element of subset. The second IF statement (line 4) checks if $word$ has run of the elements. The third IF statement (line 8) verifies if the parallel descend in $word$ and tree is possible. In the positive case, the algorithm calls $existsSubset$ with the next element of $word$ and a child of $node$ corresponding to matched symbol. Finally, if match did not succeed, current element of $word$ is skipped and $existsSubset$ is activated again in line 13.

The operation $existsSubset$ can be easily extended to find all subsets of a given $word$ in a tree $node$. After finding the subset in line 15 the subset is stored and the search continues in the same manner as before. The experimental results with the operation $getAllSubsets$(nod,word) are presented in the following section.

The operation $existsSuperset$(node,word) checks if there exists a superset of $word$ in the tree referenced by $node$. While in operation $existsSubset$ we could skip some elements from $word$, here we can do the opposite: the algorithm can skip some elements

Algorithm 4 existsSuperset($node$, $word$)

```
1:  if (not $word$.existsCurrentElement) then
2:      return true;
3:  end if
4:  found = false;
5:  from = $word$.currentElement;
6:  upto = $word$.nextElement if it exists and N otherwise;
7:  for (each $child$ of $node$ labeled $l$: $from < l \leq upto$) while $!found$ do
8:      if ($child$ is labeled $upto$) then
9:          found = existsSuperset($child$,$word$.gotoNextElement);
10:     else
11:         found = existsSuperset($child$,$word$);
12:     end if
13: end for
```

in supersets represented by $node$. Therefore, $word$ can be matched with the subset of superset from a $tree$. The operation is presented in Algorithm 4

Let us present Algorithm 4 in more detail. The first IF statement checks if we are already at the end of $word$. If so, then the parameter $word$ is covered completely with a superset from $tree$. Lines 5-6 set the lower and upper bounds of iteration. In each pass we either skip current $child$ and call $existsSuperset$ on unchanged $word$ (line 11), or, descend in parallel on both $word$ and tree in the case that we reach the upper bound ie. the next element in $word$ (line 9).

Again, the operation $existsSuperset$ can be quite easily extended to retrieve all supersets of a given $word$ in a tree $node$. However, after $word$ (parameter) is matched completely (line 2 in Algorithm 4), there remains a subtree of trailers corresponding to a set of supersets that subsume $word$. This subtree is rooted in a tree node, let say $node_k$, that corresponds to the last element of $word$. Therefore, after the $node_k$ is matched against the last element of the set in line 2, the complete subtree has to be traversed to find all supersets that go through $node$.

2. Experiments

The performance of the presented operations is analysed in four experiments. The main parameters of experiments are: the number of words in the tree, the size of the alphabet, and the maximum length of words. The parameters are named: $numTreeWord$, $alphabetSize$, and $maxSizeWord$, respectively. In every experiment we measure the *number of visited nodes necessary for an operation to terminate*.

In the first experiment, $SetTrie$ is used to store real-world data – it stores the words from English Dictionary. In the following three experiments we use artificial data – datasets and test data are randomly generated. In these experiments we analyse in detail the interrelations between one of the stated tree parameters on the number of visited nodes.

In all experiments we observe four operations presented in the previous section: $existsSubset$ (abbr. esb) and its extension $getAllSubsets$ (abbr. gsb), and $existsSuperset$ (abbr. esr) and its extension $getAllSupersets$ (abbr. gsr).

2.1. Experiment with Real-World Data

Table 1. Visited nodes for dictionary words

word length	esr	gsr	esb	gsb
2	523	169694	1	1
3	3355	103844	3	3
4	12444	64802	6	6
5	9390	34595	11	12
6	11500	22322	14	19
7	12148	17003	18	32
8	8791	10405	19	46
9	6985	7559	19	78
10	3817	3938	21	102
11	3179	3201	20	159
12	2808	2820	20	221
13	2246	2246	22	290
14	1651	1654	19	403
15	1488	1488	18	575
16	895	895	19	778
17	908	908	20	925
18	785	785	18	1137
19	489	489	22	1519
20	522	522	19	1758
21	474	474	19	2393
22	399	399	17	3044
23	362	362	17	3592
24	327	327	19	4167

Let us now present the first experiment in more detail. The number of words in test set is 224,712 which results in a tree with 570,462 nodes. The length of words are between 5 and 24 and the size of the alphabet ($alphabetSize$) is 25. The test set contains 10,000 words.

Results are presented in Table 1 and Figure 2. Since there are 10,000 words and 23 different word lengths in the test set, approximately 435 input words are of the same length. Table 1 and Figure 2 present the average number of visited nodes for each input word length (except for gsr where values below word length 6 are intentionally cut off).

Let us give some comments on the results presented in Table 1. First of all, we can see that the superset operations (esr and gsr) visit more nodes than subset operations (esb and gsb).

The number of nodes visited by esr and gsr decreases as the length of words increases. This can be explained by more constrained search in the case of longer words, while it is very easy to find supersets of shorter words and, furthermore, there are a lot of supersets of shorter words in the tree.

Since operation gsr returns all supersets (of a given set), it always visits more nodes than the operation esr. However, searching for the supersets of longer words almost always results in failure and for this reason the number of visited nodes is the same for both operations.

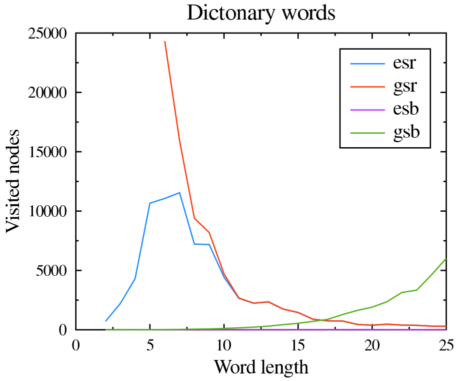


Figure 2. Number of visited nodes

The number of visited nodes for esb in the case that words have more than 5 symbols is very similar to the length of words. Below this length of words both esb and gsb visit the same number of nodes, because there were no subset words of this length in the tree and both operations visit the same nodes.

The number of visited nodes for gsb linearly increases as the word length increases. We have to visit all the nodes that are actually used for the representation of all subsets of a given parameter set.

2.2. Experiments with Artificial Data

In $experiment1$ we observe the influence of changing the maximal length of word to the performance of all four operations. We created four trees with $alphabetSize$ 30 and $numTreeWord$ 50,000. $maxSizeWord$ is different in each tree: 20, 40, 60 and 80, for tree1, tree2, tree3 and tree4, respectively. The length of word in each tree is evenly distributed between the minimal and maximal word size. The number of nodes in the trees are: 332,182, 753,074, 1,180,922 and 1,604,698. The test set contains 10,000 words.

Figure 3 shows the performance of all four operations on all four trees. The performance of superset operations is affected more by the change of the word length than the subset operations.

With an even distribution of data in all four trees, esr visits most nodes for input word lengths that are about half of the size of $maxSizeWord$ (as opposed to dictionary data where it visits most nodes for word lengths approximately one fifth of $maxSizeWord$). For word lengths equal to $maxSizeWord$ the number of visited nodes is roughly the same for all trees, but that number increases slightly as the word length increases.

esb operation visits fewer than 10 nodes most of the time, but for $tree3$ it goes up to 44 which is still a very low number. The experiment was repeated multiple (about 10) times, and in every run the operation “jumped up” in a different tree. As seen later in $experiment2$, it seems that $numTreeWord$ 50 is just on the edge of the value where esb stays constantly below 10 visited nodes. It is safe to say that the change in $maxSizeWord$ has no major effect on $existsSubSet$ operation.

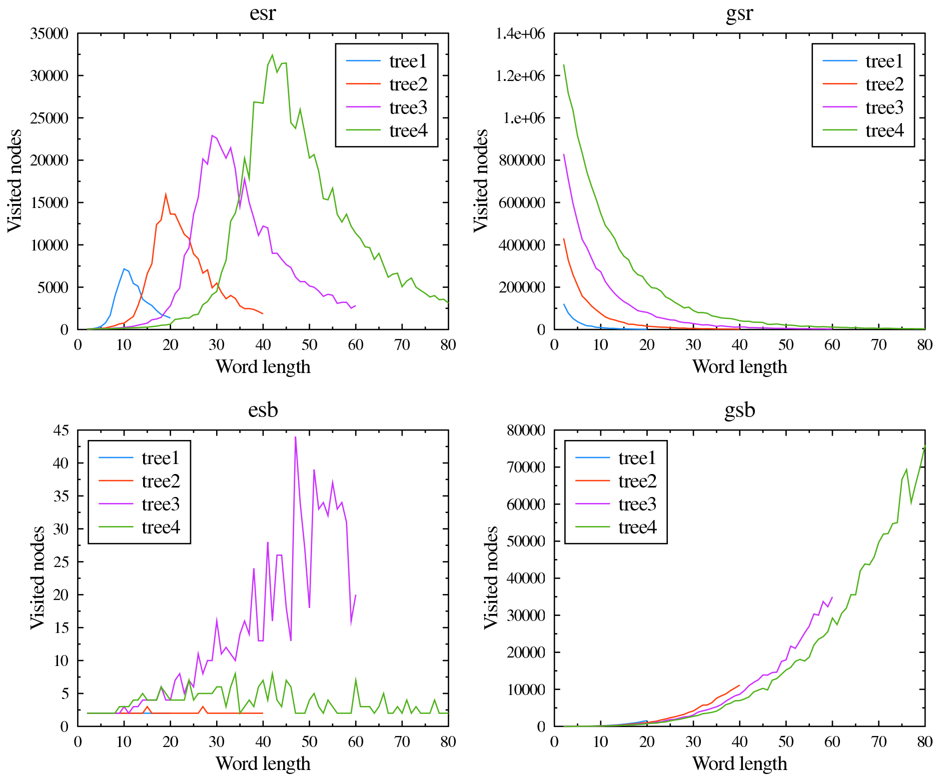


Figure 3. Experiment 1 – increasing $maxSizeWord$

In contrast to gsr, gsb visits less nodes for the same input word length in trees with greater $maxSizeWord$, but the change is minimal. For example for word length 35 in $tree2$ ($maxSizeWord$ 40) gsb visits 7,606 nodes, in $tree3$ ($maxSizeWord$ 60) it visits 5,300 nodes and in $tree4$ ($maxSizeWord$ 80) it visits 4,126 nodes.

In $experiment2$ we are interested about how a change in the number of words in the tree affects the operations. Ten trees are created all with $alphabetSize$ 30 and $maxSizeWord$ 30. $numTreeWord$ is increased in each tree by 10,000 words: $tree1$ has 10,000 words, and $tree10$ has 100,000 words. The number of nodes in the trees (from $tree1$ to $tree10$) are: 115,780, 225,820, 331,626, 437,966, 541,601, 644,585, 746,801, 846,388, 946,493 and 1,047,192. The test set contains 5,000 words.

Figure 4 shows the number of visited nodes for each operation on four trees: $tree1$, $tree4$, $tree7$ and $tree10$ (only every third tree is shown to reduce clutter). When increasing $numTreeWord$ the number of visited nodes increases for esr, gsr and gsb operations. esb is least affected by the increased number of words in the tree. In contrast to the other three operations, the number of visited nodes decreases when $numTreeWord$ increases.

For input word lengths around half the value of $maxSizeWord$ (between 13 and 17) the number of visited nodes for esr increases with the increase of the number of words in the tree. For input word lengths up to 10, the difference between trees is minimal.

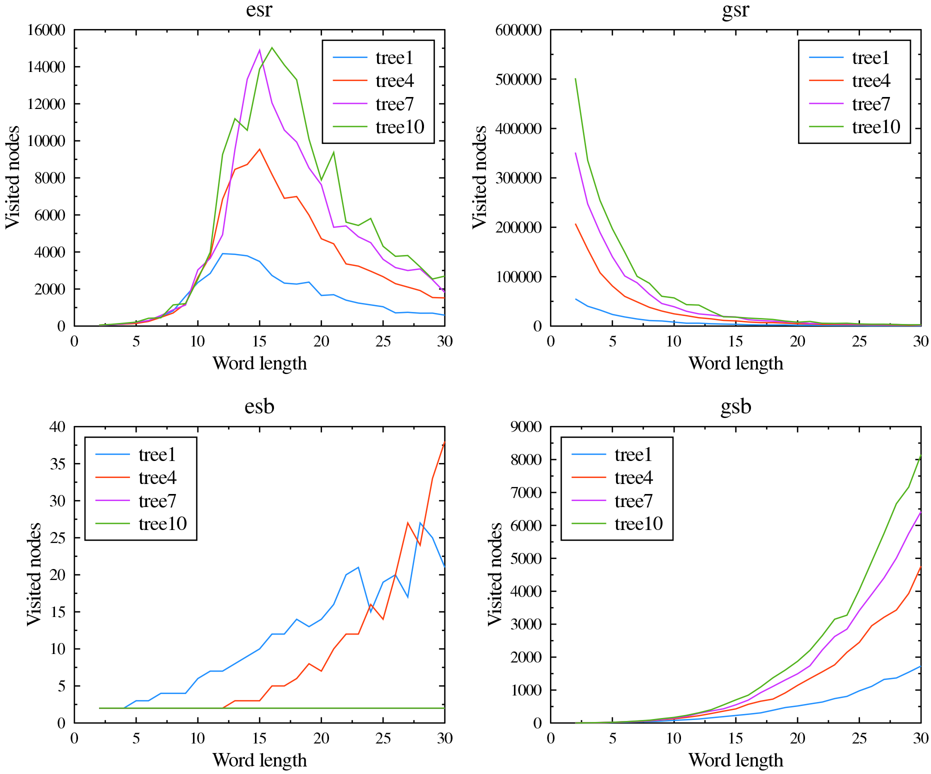


Figure 4. Experiment 2 – increasing $numTreeWord$

After word lengths about 20 the difference in the number of visited nodes between trees starts to decline. Also, trees 7 to 10 have very similar results. It seems that after a certain number of words in the tree the operation "calms down".

The increased number of words in the tree affects the gsr operation mostly in the first quarter of $maxSizeWord$. The longer the input word, the lesser the difference between trees. Still, this operation is the most affected by the change of $numTreeWord$. The average number of visited nodes for all input word lengths in tree1 is 8,907 and in tree10 it is 68,661. Due to the nature of the operation, this behavior is expected. The more words there are in the tree, the more supersets can be found for an input word.

As already noted above, when the number of words in the tree increases the number of visited nodes for esb decreases. After a certain number of words, in our case this was around 50,000, the operation terminates at a minimum possible visits of nodes for any word length. The increase of $numTreeWord$ seems to "push down" the operation from left to right. This can be seen in figure 4 by comparing $tree1$ and $tree4$. In $tree1$ the operation visits more then 10 after word length 15, and in $tree4$ it visits more than 10 nodes after word length 23. Overall the number of visited nodes is always very low.

The chart of gsb operation looks like a mirrored chart of gsr. The increased number of words in the tree has more effect on input word lengths where the operation visits more nodes (longer words). Below word length 15 the difference between trees is in the range of 100 visited nodes. At word length 30 gsb visits 1,729 nodes in $tree1$ and 8,150 nodes in $tree10$. The explanation in for the increased number of visited nodes is similar as for gsr operation: the longer the word, the more subsets it can have, the more words in the tree, the more words with possible subsets there are.

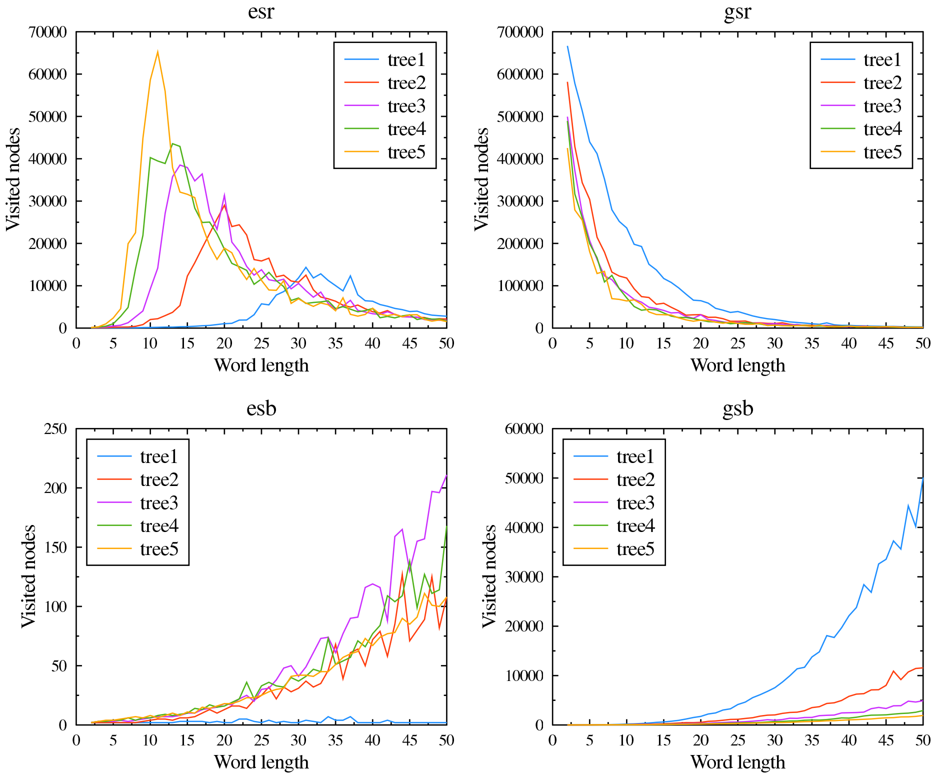


Figure 5. Experiment 3 – increasing $alphabetSize$

In $experiment3$ we are interested about how a change in the alphabet size affects the operations. Five trees are created with $maxSizeWord$ 50 and $numTreeWord$ 50,000. $alphabetSize$ is 20, 40, 60, 80 and 100, for $tree1$, $tree2$, $tree3$, $tree4$ and $tree5$, respectively. The number of nodes in the trees are: 869,373, 1,011,369, 1,069,615, 1,102,827 and 1,118,492. The test set contains 5,000 words.

When increasing $alphabetSize$ the tree becomes sparser–the number of child nodes of a node is larger, but the number of nodes in all five trees is roughly the same. For gsr and more notably gsb operation, visit less nodes for the same input word length: the average number of visited nodes decreased when $alphabetSize$ increases. The esr operation on the other hand visits more nodes in trees with larger alphabetSize.

The number of visited nodes of esr increases with the increase of $alphabetSize$. This is because it is harder to find supersets of given words, when the number of symbols that make up words is larger. The effect is greater on word lengths below half $maxSizeWord$. The number of visited nodes starts decreasing rapidly after a certain word length. At this point the operation does not find any supersets and it returns false.

gsr is not affected much by the change of $alphabetSize$. The greatest change happens when increasing $alphabetSize$ over 20 ($tree1$). The number of visited nodes in trees 2 to 5 is almost the same, but it does decrease with every increase of $alphabetSize$.

In $tree1$ esb visits on average 3 nodes. When we increase $alphabetSize$ the number of visited nodes also increases, but as in gsr the difference between trees 2 to 5 is small.

The change of $alphabetSize$ has a greater effect on longer input words for the gsr operation. The number of visited nodes decreased when $alphabetSize$ increased. Here again the biggest change is when going over $alphabetSize$ 20. With every next increase, the difference in the number of visited nodes is smaller.

3. Related work

The initial implementation of $SetTrie$ was in the context of a datamining tool $fdep$ which is used for the induction of functional dependencies from relations [8,9]. $SetTrie$ serves there for storing and retrieving hypotheses that basically correspond to *sets*.

The data structure we propose is similar to trie [6,7]. Since we are not storing sequences but *sets* we can exploit the fact that the order in sets is not important. Therefore, we can take advantage of this to use syntactical order of elements of sets and obtain additional functionality of tries.

Sets are among important data modeling constructs in object-relational and object-oriented database systems. *Set-valued attributes* are used for the representation of properties that range over sets of atomic values or objects. Database community has shown significant interest in indexing structures that can be used as access paths for querying set-valued attributes [10,5,3,11,12].

Set containment queries were studied in the frame of different index structures. Helmer and Moercotte investigated four index structures for querying set-valued attributes of low cardinality [3]. All four index structures are based on conventional techniques: signatures and inverted files. Index structures compared are: sequential signature files, signature trees, extendable signature hashing, and B-tree based implementation of inverted lists. Inverted file index showed best performance over other data structures in most operations.

Zhang et al. [12] investigated two alternatives for the implementation of containment queries: a) separate IR engine based on inverted lists and b) native tables of RDBMS. They have shown that while RDBMS are poorly suited for containment queries they can outperform inverted list engine in some conditions. Furthermore, they have shown that with some modifications RDBMS can support containment queries much more efficiently.

Another approach to the efficient implementation of set containment queries is the use of signature-based structures. Tousidou et al. [11] combine the advantages of two access paths: linear hashing and tree-structured methods. They show through the empirical analysis that S-tree with linear hash partitioning is efficient data structure for subset and superset queries.

From the other perspective, our problem is similar to searching substrings in strings for which $tries$ and *Suffix trees* can be used. Firstly, Rivest examines [6] the problem of partial matching with the use of hash functions and $trie$ trees. He presents an algorithm for partial match queries using $tries$. However, he does not exploit the ordering of indices that can only be done in the case that *sets* are stored in tries.

Baeza-Yates and Gonnet present an algorithm [1] for searching regular expressions using $Patricia$ trees as the logical model for the index. They simulate a finite automata over a binary Particia tree of words. The result of a regular expression query is a superset or subset of the search parameter.

Finally, Charikar et al. [2] present two algorithms to deal with a subset query problem. The purpose of their algorithms is similar to $existsSuperSet$ operation. They extend their results to a more general problem of orthogonal range searching, and other problems. They propose a solution for "containment query problem" which is similar to our 2. query problem introduced in Introduction.

4. Conclusions

The paper presents a data structure $SetTrie$ that can be used for efficient storage and retrieval of subsets or supersets of a given $word$. The performance of $SetTrie$ is shown to be efficient enough for manipulating sets of sets in practical applications.

Enumeration of subsets of a given universal set U is very common in *machine learning* [4] algorithms that search hypotheses space ordered in a lattice. Often we have to see if a given set, a subset or a superset has already been considered by the algorithm. Such problems include discovery of association rules, functional dependencies as well as some forms of propositional logic.

Finally, the initial experiments have been done to investigate if $SetTrie$ can be employed for searching substrings and superstrings in texts. Fur this purpose the data structure $SetTrie$ has to be augmented with the references to the position of words in text. While the data structure is relatively large "index tree", it may still be useful because of the efficient search.

References

[1] Baeza-Yates, R., Gonnet, G.: Fast text searching for regular expressions or automation searching on tries. *Journal of ACM* **43**(6) (1996), 915–936.

[2] Charikar, M., Indyk, P., Panigrahy, R.: New algorithms for subset query, partial match, orthogonal range searching and related problems. In: *Proc. 29th International Colloquium on Algorithms, Logic, and Programming*, LNCS **2380** (2002), 451–462.

[3] Helmer, S., Moerkotte, G.: A performance study of four index structures for set-valued attributes of low cardinality. *The VLDB Journal – The International Journal on Very Large Data Bases* **12**(3) (2003), 244–261.

[4] Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowl-edge discovery. *Data Mining and Knowledge Discovery Journal* **1**(3) (1997), 241–258.

[5] Melnik, S., Garcia-Molina, H.: Adaptive algorithms for set containment joins. *ACM Transactions on Database Systems* **28**(2) (2003), 1–38.

[6] Rivest, R.: Partial-match retrieval algorithms. *SIAM Journal on Computing* **5**(1) (1976).

[7] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: *Introduction to Algorithms*, Second Edition, MIT Press, 2001.

[8] Savnik, I., Flach, P.A.:, Bottom-up Induction of Functional Dependencies from Relations. In: *Proc. of KDD'93 Workshop: Knowledge Discovery from Databases*, AAAI Press, Washington (1993), 174–185.

[9] Flach, P.A., Savnik, I.: Database dependency discovery: a machine learning approach, *AI Communications* **12**(3) (1999), 139–160.

[10] Terrovitis, M., Passas, S., Vassiliadis, P., Sellis, T.: A Combination of trie-trees and inverted files for the indexing of set-valued attributes. In: *Proc. of ACM International Conference on Information and Knowledge Management* (2006).

[11] Tousidou, E., Bozanis, P., Manolopoulos, Y.: Signature-based structures for objects with set-valued attributes. *Information Systems* **27** (2002), 93–121.

[12] Zhang, C., Naughton, J., DeWitt, D., Luo, Q., Lohman, G.: On supporting containment queries in relational database management systems. In: *ACM SIGMOD* (2001).

Data Mining Applications in Healthcare: Research vs Practice

Olegas NIAKŠU and Olga KURASOVA
Vilnius University, Institute of Mathematics and Informatics,
Akademijos str. 4, LT-08663, Vilnius, Lithuania
E-mail: Olegas.Niaksu@mii.vu.lt , Olga.Kurasova@mii.vu.lt

Abstract. The paper interrogates the commonly accepted belief that data mining is widely used in medicine [8] by comparing academic advances with practical achievements in the field. The paper discusses practical usage and potential gains of data mining in healthcare facilities along with the growing number of publications indicating increasing interest to the topic in the scientific society. In order to evaluate the practical use of data mining in healthcare, a survey of tertiary hospitals in 5 countries has been conducted. The countries from diverse economic development regions were selected to cover 7 tertiary hospitals with unlike economic potential. Quantitative analysis of publications in the area of data mining applications in healthcare was made in the period of the last 8 years.

Keywords. Data mining applications, medical information systems, medical informatics

Introduction

The healthcare domain is known for its ontological complexity and variety of medical data standards and variable data quality [3, 4, 5]. Adding to this privacy consideration, making an effective and practically usable medical knowledge discovery is an open subject for the last decades. Modern clinical practices also undertake transformation not only in diagnosis, and treatment methods, but also in understanding of health and illness concepts [18].

Though data mining (DM) methods and tools have been applied in different domains already for more than 40 years, their applications in healthcare are relatively young. R. D. Wilson et al. [17] have started to classify and collect medical publications where knowledge discovery and DM techniques were applied or researched from 1966 till 2002. According to their study results "…some authors refer to DM as the process of acquiring information, whereas others refer to DM as utilization of statistical techniques within the knowledge discovery process." In fact, this mix of different concepts makes research complicated and less reliable. Therefore we decided to complement typical detailed analyses of scientific and commercial publications with surveying of large healthcare facilities, which conduct scientific and commercial research studies.

Aiming to avoid misinterpretations, the concept of data mining was defined and explained to the survey participants as follows: "Data mining, as part of knowledge discovery process, is a set of data analysis methods using statistical methods and heuristics, which are used for prediction, classification, clustering tasks or finding hidden patterns and correlations in raw data". Following examples of typical DM usage in healthcare were provided: "patient diagnostics, prediction of patient condition, prediction of post-operational complications…".

Gathering the information from hospitals allows us to put academic effort and practical usage side by side and conclude on actual DM usage, and to understand if there is a gap between data analysis experts' community and healthcare practitioners and scientists.

In this paper, we intend to combine the quantified results of publication search, which contained details of DM applications in healthcare with the results of tertiary[1] hospitals' survey on the practical DM usage. The outcome of the combination of these different sources should help us formulate a hypothesis for a further more specific and larger scale survey on the magnitude of actual DM applications in the healthcare.

Starting from the 21st century many countries have chosen e-Health as a priority national program, which in essence proposes to benefit from the standardization, aggregation of patient's clinical information and healthcare services rendered by providing instant access to that information to healthcare professionals as well as patients themselves [6, 12, 19]. According to the report [20] from the National Center for Health Statistics of USA, adoption of Electronic Health Record (EHR) in the USA as the most prominent medical information system is shown in Figure 1. It illustrates a linearly raising amount of non-sparse, but continual data reflecting patients' clinical continuity together with the treatment which took place and medication being used.

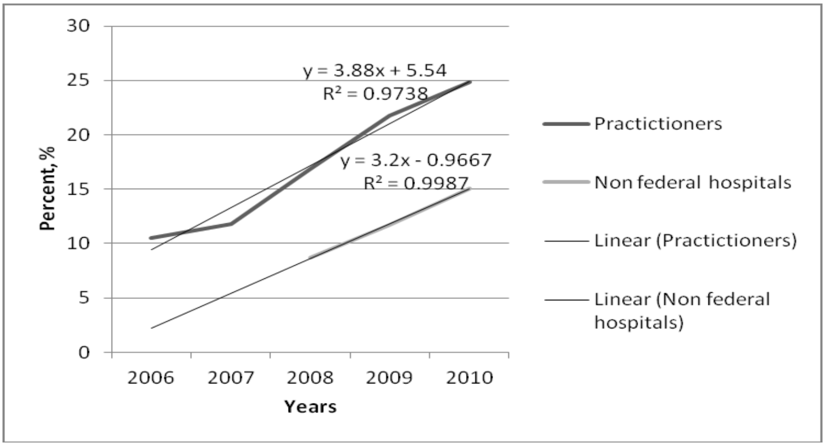


Fig. 1 EHR adoption in the USA

According to strategic plans of the EU member states, the USA and of many other nations from all continents, a considerable amount of investments is allocated to enable the global computerization of healthcare data. Taking a linear progression would mean that in 10 years all new medical encounters will be thoroughly digitalized in all developed countries. The exponential growth is doubtful mostly because of lack of governance structures, data protection and patient privacy issues and resistance from inside of medical community [12]. But even considering a conservative scenario, it is

[1] Tertiary hospital – a major hospital, providing wide range of high level specialized medical services. Commonly tertiary hospitals are university hospitals combining medical and academic activities.

becoming obvious that, for the first time in the history, research community is going to get a full set of a person's medical history from the birthdate till he or she passes away. And that is not for a small specific group limited by longitudinal research study, but the whole regions, nations, countries and even continents. This anticipated scenario forecasts tremendous potential for machine learning and in particular for DM applications in healthcare.

1. Scope of Analysis

Thomson Reuters Web of Science [26], Google Scholar [22] and PubMed [25] databases were used to analyze the number and distribution of scientific publications related to DM in medicine in the last decade.

Tertiary hospitals were selected as a primary source for our survey. The main reason is that typically, tertiary hospitals are in the first line of healthcare institutions that implement clinical software systems, enabling to collect clinical and demographical patient data needed for DM applications. Historically hospital information systems developed starting from the exotic show cases of the economically well-established communities to the standardized practice of handling clinical data and workflows since mid-nineties in developed countries and from the first decade of the XXI century in the developing countries and emerging markets.

2. Scientific Relevance and Development

2.1. PubMed Database

PubMed database is comprised of more than 21 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed is operated by National Healthcare Library of U.S. and indexes all publications classifying its content with the help of MESH structured vocabulary [24]. Using MESH vocabulary terms as a search parameter in PubMed database guaranties that not only search wording matching publications will be found, but also its matching synonymic wording or previously used terms. MESH term, classified as MESH heading "data mining" is mapped to other similar concepts like "text mining". "data mining" term was appended to the vocabulary only in 2010 and the former terms e.g. "Information Storage and Retrieval", previously used for the same or similar and related concepts, are mapped to the latest one. A simple search criterion "data mining" was used to retrieve a number of publications and books within the medical domain with assigned MESH heading "data mining". The first publication is dated 1984, however the second one appears only after 10-year interval in 1994. This search resulted in 3077 publications.

2.2. Thomson Reuters Web of Science Database

Web of Science has been providing access to more than 12,000 journals in all subject area. It also includes citations to conference proceedings. The advanced search filter allows the use of logical operations, search restricted to the selected subject areas, and the search scope (title of the publication or whole text). The following constraints have been chosen for our analysis purposes:

> (TS=(data mining) AND TS=(medic* OR clinical OR healthcare)) AND Document Types=(Article OR Abstract of Published Item OR Proceedings Paper)
>
> Refined by: [excluding] Web of Science Categories=(OPERATIONS RESEARCH MANAGEMENT SCIENCE OR TELECOMMUNICATIONS)
>
> Timespan=1996-2012. Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH. Lemmatization=On

This search resulted in 2272 publications.

2.3. Google Scholar

Google Scholar provides a scholarly literature search service across many disciplines and sources, including theses, books, abstracts and articles. However it is not limited to scientific publications only. Google Scholar indexes content items published since 1993. Google search filter allows the use of logical operations AND, OR, NOT, a restricted search only in the selected subject areas, search scope (title of the publication or whole text). The following constraints have been chosen for our analysis purpose:

> Search in the title: "data mining" AND (medical OR clinical OR medicine OR healthcare)

This search resulted in 478 publications.

The choice of searching in the whole article text was rejected due to a serious flaw: a huge amount of DM centric publications have keywords “medicine” or “healthcare” in the text with a purpose to illustrate DM usage. But this proved to be insufficient to indicate that a publication is focused on DM applications in medicine.

Google indicated the number of publications in the selected period of time approximately. And according to our observations, precision is increasing with a larger quantity of the relevant content items found.

2.4. Results

Distribution of publications found in Web of Knowledge, Google Scholar and PubMed databases starting from 1997 to 2011 is shown in Figure 2.

We can see mostly a linear growth in all the cases, with different line slope: in Google Scholar case it is a very symbolic growth m ≈ 3,38, more significant in Web of Knowledge m ≈ 18,28 and finally in PubMed database m ≈ 30,5. As seen from the explanation of the queries searched in two databases, the results are not directly comparable and are shown here to illustrate a constant raising interest of the academic society in the topic of DM applications in medical domain.

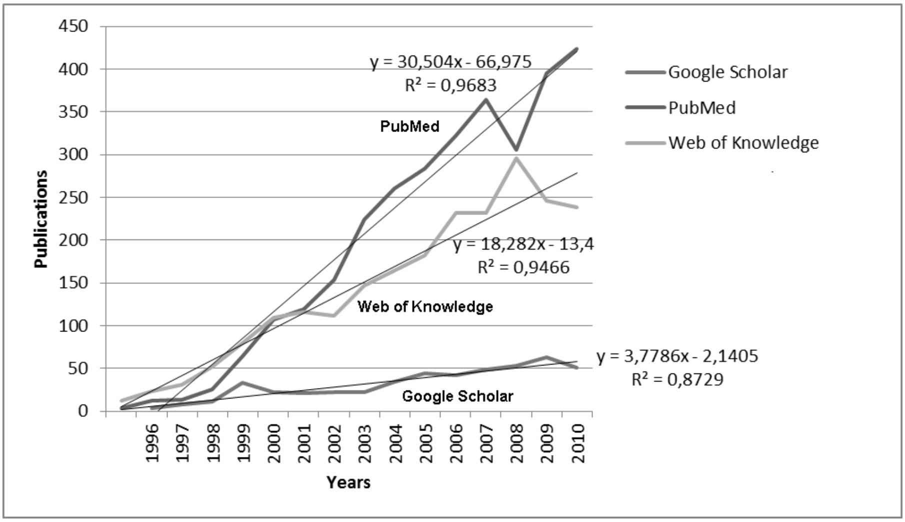


Fig. 2. Trend lines of DM applications in medicine related publications

2.5. Public Interest in the Topic of DM

Though it is difficult to estimate exact numbers, however we can get a fair understanding using a publicly available tool Google Trends [23], which analyzes all search queries executed in Google search engine worldwide. Google Trends service has been collecting and mining data since 2004, providing time-series analysis, reflecting the overall actuality of different search topics or “trends” as Google names it. The dimensions of geographical location, source and language are taken into consideration. We can get a better understanding by comparing the actuality of the term to other disciplines, like *artificial intelligence*, *machine learning*, or a broader concept like *computer science*. This type of analysis provides a unique source of unified data that combines search queries in different languages from different world locations.

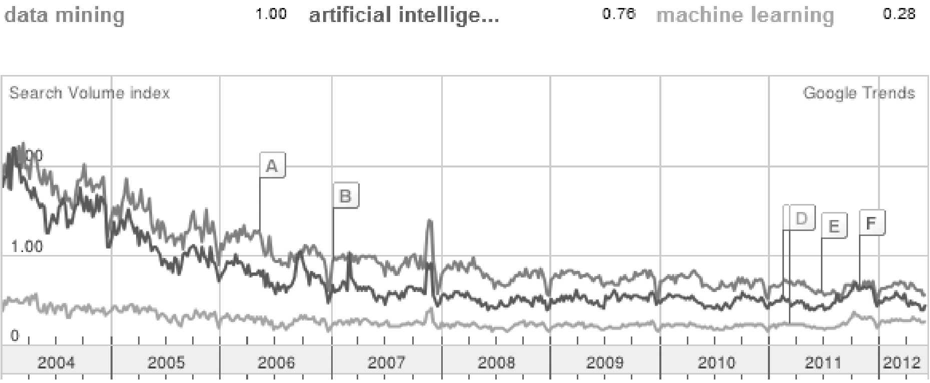


Fig. 3. Google Trends in DM, artificial intelligence and machine learning

Google trends is not providing absolute values on vertical axis, instead, chart's data is scaled to the average search traffic for ”data mining” term (represented as 1.0) during the time period from 2004 till 2012.

As shown in Figure 3, there is a general correlation among all the concepts analyzed: *artificial intelligence, machine learning* and *computer science*. Addition to the analysis such disciplines as “mathematics” or “physics” will show the same correlation. A little decrease of the general public interest is noticeable in scientific or, we should say, scholarly topics over years.

Also, looking at Figure 4, which shows the trends in more detail for 2011, we can see a decrease of interest in all topics during summer months, which is most probably a direct indication of summer holidays in the academic society.

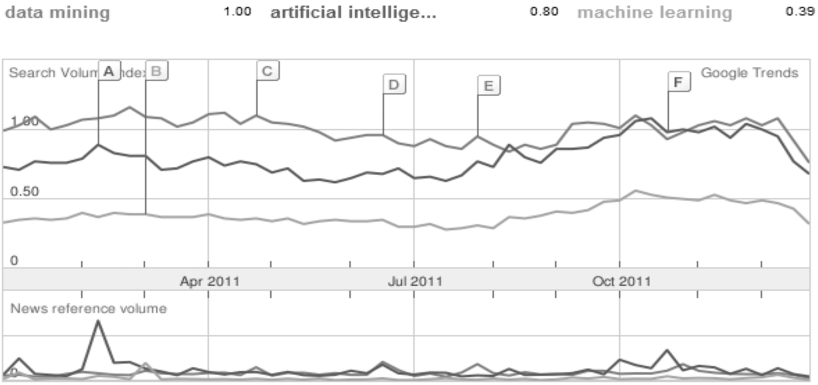


Fig. 4. DM Google trends during 1 year

Figure 3 illustrates that at least during last 8 years, DM topic has a more or less stable interest among the general public, but we assume this interest to be mostly related to scholarly activities. Another interesting outcome of Google Trends analyses was distribution of top 10 geographical locations of the searching query sources (Figure 5).

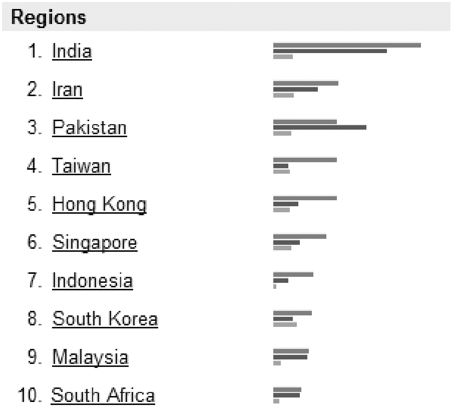


Fig. 5. DM Google Trends by regions

This distribution can be interpreted in different ways. However, it offers an additional perspective in understanding, which regions will be more active in the field in the nearest future.

Summarizing Google Trends results, we can conclude that the peak popularity of "DM" and "artificial intelligence" concepts finished by 2007 and afterwards it remains more or less stable.

Trying to find correlations between DM actuality in the academic world and general public would not be correct because the provided trends reflect different enquiries: scientific result creation versus generic interest in the topic without any obligation or intention to create any sort of result out of it.

3. Surveying DM Applications in Healthcare Facilities

As shown above, the volume of medical related DM research increases from year to year. Naturally, one can suppose that DM usage penetration is increasing accordingly. However, a data analyst working in the field will agree that a large number of research studies remains academic and has no clinical follow up and even rarely goes beyond the institutions which were directly involved in the research. And this already generates reasonable questioning and doubts on the rational and measurable outcome of the research effort. Undoubtedly, a number of proven DM niche applications counts award winning successes, like in radiology imaging or genetics analysis. But this cannot be said about hundreds of specific clinical DM research. Up till our research date, we could not find an example of a systematic approach in an attempt to understand history or the current situation of DM utilization by healthcare institutions. And that can be considered as a blurring factor, preventing the scientific society from concentration on correct ways of the knowledge discovery process tailored for healthcare, which would score the maximum benefit for the clinicians as end users of DM tools and methods and finally patients as beneficiaries and final added value recipients.

Due to the fact that the healthcare sector is very diverse and its entities as well as actors have different objectives and fields of activities, they, employ different methods and tools in their operations [1, 2, 11, 13, 14, 15, 16]. Therefore it was initially agreed to define the scope of this research as DM applications in the healthcare providers' institutions. However, for this scope the statistically valid representation would require a significant number of different type of institutions like General Practitioners offices, private or public clinics, local and specialized hospitals, regional and, finally, tertiary hospitals. The initial experience of interviewing healthcare institutions suggested that the highest probability of DM usage will score in tertiary hospitals, which have tight relations with the academic society and participate in different sorts of scientific and commercial research on a regular basis. That does not lead to the conclusion that DM applications in smaller institutions do not occur, but focusing on tertiary hospitals allowed us to estimate the upper range of DM penetration into healthcare providers sector. The next important surveying scope constraint is geographical spread of healthcare institutions (HCI). Setting the initial research objective, to understand the practical usage of DM techniques and tools in HCI across the globe puts a very ambitious but unrealistic target to survey thousands of healthcare facilities. It has been decided, that for a limited resources study we should select tertiary hospitals at least from different zones of economic development, having a different magnitude of electronically available patient related data for further analyses. Therefore it was extremely important to select well financed hospitals from the leading economies countries as well as relatively modestly financed hospitals from the developing countries. Hospitals from the following countries participated in the survey: South African Republic, Lithuania, Switzerland, Albania, and Germany. This survey cannot be treated as final as we plan to continue gathering information in upcoming years; however it reveals clear patterns which lead us to concrete conclusions and summarizations that might be useful for both communities of data analysts and clinicians.

3.1. Preparation and Conducting the Survey

Already in early stages it has become obvious, that there is a huge gap in understanding of DM concept by its intended end users - clinicians. Typically hospital's IT department has knowledge and is able to describe how DM is used in the hospital. On the contrary, medical personnel are usually minimally informed or knowledgeable about what exactly DM is and more specifically, how it is used in the hospital. Accordingly, we are in a situation when we cannot ignore either the first class of respondents or the second one. And it was important to get both types of answers for later analysis. Afterwards, we have analyzed the answers classifying both respondents' classes separately and summing them up together. Taking this diverse interviewing audience into consideration, questions were formulated in a comprehensible way for a broader range of respondents with a medical or IT background. See the summarized questions below.

Usage of statistical data analyses, DM and clinical decision support systems:

1. Have you heard about practical applications of DM in medicine?
2. Do you know any research projects in your hospital using DM methods?
3. Have you or your colleagues been involved in DM research project, aiming to identify new patterns or finding new rules for patient diagnostics, prediction of treatment results or other. If yes, please provide a brief summary of research aim and the results.
4. If DM methods have been used, was your experience successful? Please comment
5. Has the clinical decision support IT system been used in your hospital?
6. Please specify which clinical specialties could benefit by using DM methods on collected patient clinical data in your hospital (choose from the list)
7. What type of clinical research your hospital is involved in?
8. Are you or your colleagues potentially interested in the benefits which DM could provide to you?

Availability of Electronic Patient Data for Research:

9. How many years have the patient data been collected in IT systems in your organization?
10. Please specify what clinical patient information is stored in IT systems (HIS, EHR, EMR, RIS, etc.). Select from the list: Observations, Lab results, Radiology reports, Anamnesis, Surgery reports, Discharge summary, Visit summary, Nursing data (vitals), Medication used (for inpatients).
11. Mark medical IT systems used in your organization. Select from the list: EMR / EPR, HIS, RIS/PACS, LIS, Specific clinical information systems, Emergency IS, OP clinic information system, Blood bank information system, Clinical decision support system, Pathology information system.

12. Specify what standard nomenclature is used in your organization (e. g. ICD9, ICD10, SNOMED-CD, LOINC). Select from the list: Patient diagnosis, Pathologic diagnosis, Procedure coding, Laboratory coding.

Interest in DM:

13. Are you interested in international clinical DM research projects?
14. Specify the clinical specialty or problem you are interested in.

3.2. Method of Survey

The survey was conducted according to methodical guidelines of the Centre for Health Promotion of University of Toronto [21]. A call for survey was openly published in the eHealth news portal eHealthServer.com [27]. The survey was prepared in an online questionnaire and offline forms. The need for an offline version was pointed by some institutions with a limited or no internet access. Hospitals were asked that at least 2 respondents from each institution should fill the questionnaire; a person in charge for medical services, e.g. medical superintendent, director of medicine, head of the clinical department and a person in charge for Information Technology e. g. chief of the IT department. In parallel direct enquiries were sent to the officials of hospitals in 8 countries. Complete interviewing took five months instead of two months planned due to very little or no reaction from the respondents, especially from medical representatives.

The survey's questions allowed crosschecking correctness of the information provided. E. g. question #4 asks explicitly if DM tools are used and question #5 asks if a clinical decisions support system is in use. Typically a clinical decision support system would incorporate a few DM algorithms as well as statistics.

The aim of questions in the section "Usage of statistical data analyses, DM and clinical decision support systems" is to clarify the eligibility of the institution for DM, awareness of DM concept and known applications of DM.

Questions in the section "Availability of Electronic Patient Data for Research" help to figure out the potential of DM in the institution, based on the amount of electronically available data, the medical information system, and standardized medical nomenclature being used.

The aim of questions in the section "Interest in DM" is to define what the interest of the respondents is in possible future DM research projects.

4. Analysis of Survey Data

Out of 14 respondents 12 have confirmed that they had heard about practical applications of DM. However, after the quality validation and answer crosschecking, only 9 positive answers could be qualified. But even out of the remaining 9 respondents with positive answers only 4 are familiar with practical examples of such usage, making up 29% of the whole. Another aspect provided by data validation, is that the majority of medical respondents would have no information about DM research initiatives and applications in their own facilities. It is difficult to specify the overall level of awareness in terms of this survey; however, the selected method of surveying 2 and more representatives from each facility has proved that typically medical specialists, not related to the DM project in their own HCI, have no information about it.

Table 1. Summary of survey answers

	Hospitals of developing countries		Hospitals of emerging countries		Hospitals of western countries	
Summarized questions	IT rep	Clinical rep	IT rep	Clinical rep	IT rep	Clinical rep
Understanding, practical usage and interest						
Good understanding of DM concept	Yes	Differs	No	No	Yes	Differs
Awareness of practical use	No	No	No	No	Differs	No
Hands on DM applications	Differs	No	No	No	Yes	No
Interest in the topic	Yes	Yes	Yes	Yes	Yes	Yes
Clinical specialties	All	All	All	All	All	All
Availability of electronic data for research						
Number of years data is electronically captured	4-13 years		1-3 years		5-15 years	
Variety of medical information systems used to capture and operate with patient related data	patient demographics, radiology images, partly lab results, partly detailed clinical data, billing data		patient demographics, limited radiology images, partly billing data		patient demographics, radiology images, lab results, detailed clinical data, billing data	

Summarized answers, grouped by hospitals with alike economical situation are provided in 1 table. Country groups are represented as follows:

- developing countries – Albania;
- emerging countries – Lithuania and South African Republic;
- western countries - Switzerland and Germany.

Evaluating the benefits of gained DM experience, 50% of respondents, who declared a personal involvement in DM projects, were satisfied with the results achieved and 50% had a neutral opinion on the project success.

Analysis which clinical specialties have the highest potential in DM usage was not successful. Validation of answers has showed that typically all the selected clinical specialties were relevant either to the clinical profile of the respondent or to the clinical profile of the hospital. Summarizing the answers provided, we can conclude, that all clinical specialties without an exception have a potential for DM.

The interest in additional information on potential DM benefits was expressed by 86% of respondents, regardless of their initial experience with DM.

The analysis of electronic data availability for DM purposes showed us a correlation between depicted years of clinical data collection in a facility with the level of the region's economic development (Figure 6). Data collection timeframe values spread from 1 to 15 years, with the mean value 8 years and median value 4 years. In terms of medical IT systems being used, 100% of respondents have defined that hospital information system is in use; electronic medical record systems are used in 60% of facilities and radiology imaging systems in 83%.

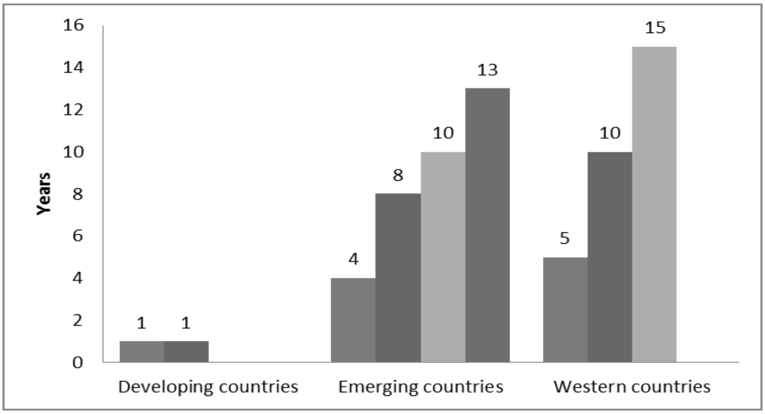


Fig. 6 Clinical patient data collected electronically in hospitals

The usage of standard terminology dictionaries varies depending on the originating country of the facility. The usage of ICD 9 and ICD 10 is very common for coding disease diagnoses. However, other nomenclatures, critical for DM applications and used to code procedure/intervention, laboratory tests, pathology diagnoses, are only partly implemented and at a different quality level.

93% of respondents expressed their will to participate in international clinical DM research projects.

5. Survey Findings

As it was presumed, understanding of DM as a concept as well as its potential depends on the background of the respondents. IT personnel of a hospital typically are well informed on the DM related research and usage inside of the hospital, scoring 100% of its surveyed IT department representatives. In addition, clinicians usually informed only if they were directly involved in such projects.

All the respondents have confirmed that they had heard about practical applications of DM in medicine. However, only 29% of respondents were able to provide any example of practical DM usage.

There is a noticeable confusion in differentiating DM and statistics concepts among healthcare professionals, and very rarely DM is treated by them as a practically valuable tool for clinical purposes.

The respondents from healthcare facilities with a relatively recent adoption of IT in the patient treatment process tend to mix statistical reporting and DM, hospital information systems, Electronic medical record systems and decision support systems.

Regardless of understanding and experience of DM, 86% of respondents expressed their interest in the DM topic and 93% would like to participate in international DM research projects as well as to be informed about utilization of DM techniques in the future.

6. Conclusions

The analysis of publications in the field of DM application in the medical domain has shown a steady growth since its accountable beginning till nowadays. The line slope of publications growth can be averaged to m ≈ 17 on the search conducted in PubMed, Web of Science, and Google Scholar databases. In the early 90'ties up to 5 publications were produced during one year and around 400 publications in 2011. We can conclude that a tremendous growth of interest and scientific advancement took place in the last decade.

On the DM value chain's side, survey revealed, that the greatest part of medical community of tertiary hospitals have either minimal or zero awareness of the DM practical usage and its potential possibilities. All the respondents from the largest university hospitals confirmed to be familiar with DM applications in healthcare, however only 29% of them were able to provide any example of practical DM usage. A huge gap in awareness and understanding of the DM potential was encountered even inside healthcare facilities splitting IT and the clinical personnel to different poles. If we interpolate these results to the smaller, less financed and less exposed to research projects healthcare providers, DM usage will be significantly lower. The survey identified a considerable potential for a further DM penetration due to an increasing amount of patient clinical data collected in HCI and interest declared by hospitals' clinical representatives: 86% of respondents expressed their interest in DM and even more would like to participate in international DM research projects.

However, the process of information digitalization in the developing countries is still in the early phases and the lack of electronically available data is a stopping factor for the spread of DM in a poor economic area.

Summarizing, we have showed that data mining perception and practical applications in healthcare is a way beyond its steady growth in the academic research field, which raises a hypothesis, that relatively a little percentage of academic research effort results in practical DM applications in healthcare, out of which we can conclude that the current interdisciplinary approach is not efficient enough. When considering the potential and benefits of knowledge discovery using DM tools in healthcare, it is clear that more attention should be paid to the domain specific problems of successful DM application in healthcare [4, 5, 10], emphasizing the usage of DM methods with self-explanatory models [7, 9, 16] in contrast to black-box methods.

Further research will be continued aiming to collect additional survey data from the USA, Middle East, Asia, and Australia to increase data representation and get more accurate results.

References

[1] R. Bellazzi and B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics* **77** (2008), 81–97.

[2] P. Berka, J. Rauch, and D. A. Zighed, *Data mining and Medical Knowledge Management– Cases and Applications*. Idea Group Inc (IGI), 2009.

[3] O. Bodenreider, Ontologies for mining biomedical data. In: *IEEE International Conference on Bioinformatics and Biomedicine*, Philadelphia, Pennsylvania, 2008.

[4] H. Chen, S. Fuller, C. Friedman, and W. Hersh, editors, *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*, Springer Science, 2005.

[5] K. J. Cios and G. W. Moore, Uniqueness of medical data mining, *Artificial Intelligence in Medicine* **26**(2002), 1–24.

[6] D. Castro, *Explaining International IT Application Leadership: Health IT*. The Information Technology @ Innovation Foundation, 2009.

[7] G. Dzemyda, O. Kurasova, and V. Medvedev. Dimension reduction and data visualization using neural networks emerging. In: I. Maglogiannis, K. Karpouzis, M. Wallace, J. Soldatos, editors, *Artificial Intelligence Applications in Computer Engineering* **160** (2007), IOS Press, 25-49.

[8] H. C. Kob and G. Tan. Data mining applications in healthcare, *Journal of Healthcare Information Management* **19**(2) (2005), 64-72.

[9] G. A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information, *The Psychological Review* **63** (1959), 81-97.

[10] P. M. Pardalos, V. L. Boginski, and A. Vazacopoulos, editors, *Data Mining in Biomedicine*, Springer Science, 2007.

[11] D. Ruben and Jr. Canlas, *Data mining in Healthcare: Current Applications and Issues*, Thesis, Carnegie Mellon University, Australia, 2009.

[12] K. A. Stroetmann, J. Artmann, and V. N. Stroetmann. *European Countries on their Journey Towards National eHealth Infrastructures*. Final European Progress Report. European Commission, DG Information Society and Media, ICT for Health Unit, 2011.

[13] W. Stühlinger, O. Hogl, H. Stoyan, and M. Müller. Intelligent data mining for medical quality management. In: *Workshop Notes of the 14th European Conference Artificial Intelligence*, 2000, 55–67.

[14] V. Špečkauskienė and A. Lukoševičius. Methodology of adaptation of data mining methods for medical decision support: case study, *Electronics and Electrical Engineering* **2**(90) (2009), 25–28.

[15] P. Treigys, V. Šaltenis, G. Dzemyda, V. Barzdžiukas, and A. Paunksnis, Automated optic nerve disc parameterization, *Informatica* **19**(3) (2008), 403-420.

[16] S. Wasan, V. Bhatnagar, and H. Kaur, The impact of data mining techniques on medical diagnostics, *Data Science Journal* **5** (2006) 119–126.

[17] A. Wilson, L. Thabane, and A. Holbrook. Application of DM techniques in pharmacovigilance, *British Journal of Clinical Pharmacology* **57**(2), (2003), 127-134.

[18] H. R. Wulff, S. A. Pedersen, and R. Rosenberg, *Philosophy of Medicine an Introduction*, Blackwell Scientific Publications, Oxford, 1990.

[19] *Healthcare Information and Management Systems Society. Electronic Health Records. A Global Perspective*. White paper. HIMSS Enterprise Systems Steering Committee and the Global Enterprise Task Force, 2010.

[20] *Division of Health Care Statistics*. NCHS Health E-Stat Report. National Center for Health Statistics of US 2011.

[21] *Workbook. Conducting Survey Research*. Centre for health Promotion of University of Toronto, 1999.

[22] *Google Scholar* – Web search engine, indexes scholarly literature. Available from: http://scholar.google.com/.

[23] *Google Trends*. Web portal. Available from: http://www.google.com/trends/.

[24] *National Library of Medicine* – MeSH. Available from: http://www.nlm.nih.gov/mesh/meshhome.html.

[25] *PubMed – Database of References and Abstracts on Life Sciences and Biomedical Topics*. Available from: http://www.ncbi.nlm.nih.gov/pubmed/.

[26] *Web of Science* – Academic Citation Index Provided by Thomson Reuters. Available from: http://apps.isiknowledge.com/

[27] *eHealth Server news portal*. Available from: http://www.ehealthserver.com/research-and-development/935-survey-on-application-of-data-mining-to-support-clinical-decisions.

On Multi-Objective Optimization Aided Visualization of Graphs Related to Business Process Diagrams

Vytautas JANČAUSKAS [a], Giedrius KAUKAS [b], Antanas ŽILINSKAS [a,1] and Julius ŽILINSKAS [a]

[a] *Institute of Mathematics and Informatics, Vilnius University, Lithuania*
[b] *ORGSOFT, Vilnius, Lithuania*

Abstract. A problem of the drawing of aesthetically looking graphs, related to business process diagrams, is considered. We model a situation where sites of flow objects of the diagram are fixed, and the sequence flow is defined. The edges of a graph, which represent the sequence flow, should be drawn aiming at an aesthetical image. The latter problem is reformulated as a multi-objective combinatorial optimization problem. The generally recognized criteria of aesthetical presentation, such as general length of lines, number of crossings, and number of bends, are considered the objectives to be minimized. Two algorithms are developed for the stated problem taking into account its specifics. The efficiency of the developed algorithms is evaluated experimentally using randomized test problems of different complexity.

Keywords. Business process diagram, optimization, modeling, orthogonal connectors, business process management

Introduction

The diagrammatic visualization is an important aid in various fields of management and engineering The aesthetic attractiveness is a natural advantage of a drawing. Moreover, according to the general opinion, aesthetical layouts are also more informative and practical [1]. On the other hand the criteria of the aesthetic attractiveness do not always guarantee the informativeness of the diagrams drawn, as it is shown by the experiments with the CASE related diagrams in [15]: "While different generic algorithms, embodying a variety of aesthetics, may produce diagrams that look attractive, a "nice" layout is unlikely to be sufficient for intuitive use". For a discussion on the graph drawing aesthetics we refer to [2], [15], [16]. Although the problem of graph drawing attracts many researchers, and plenty of publications are available, special cases of that problem frequently cannot be solved by straightforward application of the known methods and algorithms. We cite [15] again: "Few algorithms are designed for a specific domain, and there is no guarantee that the aesthetics used for generic layout algorithms will be useful for the visualization of domain-specific diagrams". In [7] aesthetical visualization of aesthetic visualization

[1] Corresponding author

of more specific graphs, business process diagrams, is considered. It is emphasized there, that layout preferences of different user groups can differ essentially, and a set of layout criteria is formulated.

In the present paper we consider a particular problem of the aesthetical drawing of special graphs which are related to business process diagrams of small-medium enterprizes (SME). The algorithms for the aesthetically pleasing visualization of edges of those special graphs are considered, where graphs model business processes, and sequence flows should be visualized assuming the flow objects fixed. Our idea is to reduce the original problem to a problem of the combinatorial multi-objective optimization. For the discussion on the synergy of optimization and visualization we refer to [20]. The developed algorithms are aimed at including into a relatively simple and not expensive software package oriented not only to the consultants of business management but also to the practitioners in SME management [13].

1. Sequence Flow Visualization as a Special Case of Graph Drawing Problem

Our interest in this graph drawing problem is motivated by a request from the developers of a software package for modeling business processes in SME [13]. The latter is oriented to managers and consultants who either design a new SME or search for the possibilities to improve an existing one. The considered business process management methodology is oriented to managers and consultants either designing a new SME or searching for the possibilities to improve an existing one. The Business Process Modeling Notation (BPMN) is accepted as a standard for drawing Business Process Diagrams (BPD) [14]. In the present paper a partial problem of drawing the BPD is considered, namely the problem of drawing the lines which represent the sequence flow for fixed flow objects and defined sequence flow. For the more general problems of constructing BPD we refer, e.g. to [8], [10].

The problem of drawing aesthetical layouts is reformulated as a combinatorial multi-objective problem where the objectives correspond to the criteria usually used to assess the aesthetic attractiveness of a BPD. The developed algorithms can be used interactively when the flow objects are placed by a human user. We are going to continue this research, and subsequently to develop an upper-level algorithm (with respect to the considered in the present paper) for the re-location of the flow objects thus improving the overall aesthetical attractiveness of the considered BPD. The algorithms considered in the present paper will be used by the upper level algorithm as an auxiliary routine for searching Pareto optimal edges for the location of vertices analyzed at upper level.

The problem of drawing the sequence flow is a special case of drawing the edges of a graph where vertices are located on a plane. Moreover, the edges as well as the location of vertices should satisfy some special restrictions. To enable the user to completely understand the information presented by the drawing it normally contains up to 30 flow objects. Therefore in the experiments below we consider graphs with the number of vertices of up to 30. The navigation by the user in BPD is aided by visualization of well perceivable sub-graphs of the considered BPD. The more detailed information concerning the specified flow objects can be extracted by telescoping. For example, a rectangle in the BPD can represent the process, the sub-process, and the task; a process can be decomposed by means of the creation of the child BPD which shows the details of the parent BPD [14]. To denote the flow objects in the considered diagrams, the shapes of three types are used: rectangles, rhombs, and circles which represent the processes, the gateways, and the events correspondingly. The shapes are located in the pool lanes. The sequence flow is represented by the lines constituted of orthogonal segments. BPD can be augmented by data objects and data flows.

In the present paper we focus on the problem of connector routing. Therefore the differences of the flow objects are ignored, and a single rectangular shape is used below to represent the flow objects. Visualization of the graphs, where vertices are drawn as rectangles connected by piecewise vertical and horizontal lines, is commonly used. As the examples, Entity Relationship and UML diagrams can be mentioned among others. Many methods and software implementations of algorithms are available for representing graphs as rectangles connected by orthogonal connectors. However, the immediate application of the available algorithms to the visualization of business processes according to the requirements of the business processes management methodology of [13] is difficult. Of course, basic requirements to the connectors are common for the methods developed to similar problems, and therefore the ideas of known methods were useful in solving our problem.

After a discussion on the general prerequisites related to the orthogonal connectors representing the sequence flow in a business process model, more specific requirements can be stated. The rectangle shapes are located in the centers of a rectangular mesh. The segments of the orthogonal connectors can stretch along borders of the pool lines and in the passages which are orthogonal to these lines, and interpose between the cells of the mesh. An example presented in the Figure 1 illustrates permissible ways for the routing: a connector should join the neighboring vertices represented by small grey circles.

An edge of the graph which models a BPD can be represented by many orthogonal connectors. The abstract criterion of aesthetical image of a BPD depending on the connector can be decomposed into several criteria, and some of these criteria can be evaluated quantitatively. We consider three criteria which seem essential and can be relatively simply evaluated: the total length of connectors, the number of bends, and the number of crossings. All criteria should be minimized. The problem of drawing an aesthetically looking sequence flow is reduced to a multi-objective optimization problem. To the best knowledge of the authors the re-formulation of the initial problem as a multi-objective optimization problem is original although various versions of single-objective optimization problems have been investigated. In the present paper we consider all criteria equally important. However, the relative importance of the criteria can depend on the users of the supposed software package. The relevance of the considered quantitative criteria to the criterion of the subjective perception, and the relative importance of the quantitative criteria are analyzed in [21] by means of a psychological experiment.

2. A Brief Overview of Available Algorithms

The problem of drawing graphs, with the rectangular images for vertices, and with edges composed of pieces of vertical and horizontal lines, is considered in many papers. Depending on the application area in question, the algorithms must satisfy different requirements. Some algorithms, efficient from the point of view of general complexity theory, are described by [19], and [11]. A comprehensive review of algorithms oriented to the

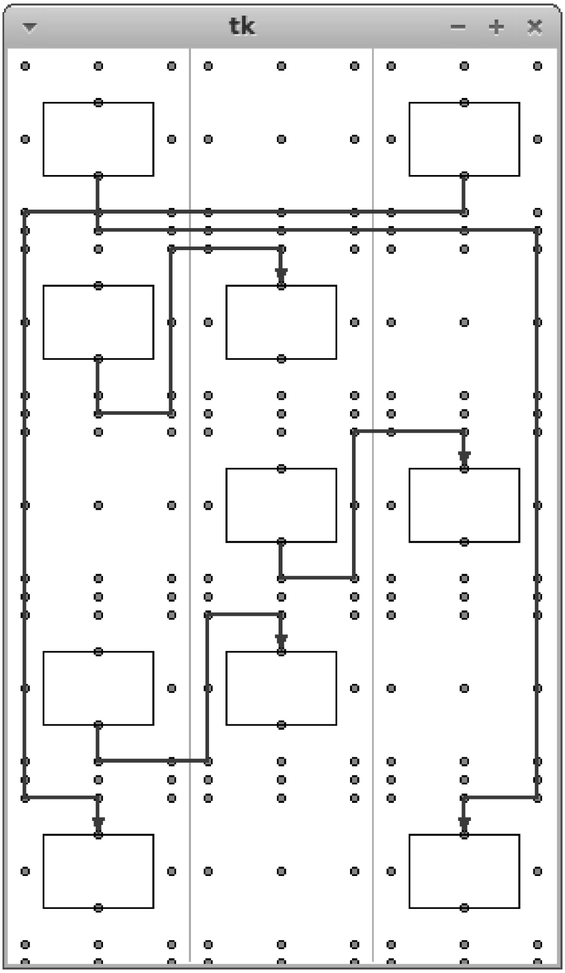


Figure 1. A graphical illustration of a solution to the routing problem

routing of paths for nets on the chip layout to interconnect the pins on the circuit blocks or pads at the chip boundary is presented in [3]. The general purpose routing algorithms are classified in three groups, namely, the maze, line-search, and A*-search groups. Since those algorithms are based on general graph-searching techniques they can be adapted to the specific requirements of both global and detailed routing problems. Different versions of those algorithms are proposed and investigated with the focus on the asymptotic complexity estimates and on the application in the chip design. From the point of view of the BPD drawing, the criteria of aesthetics prevail the criteria important in technological applications emphasized in [3]. In a recent paper by Wybrow et al [18] a brief review of the available algorithms and software, for the construction of orthogonal connectors, is presented from the point of view of the requirements similar to those stated in the previous section. The experimental testing performed by these authors has shown that some of the available software packages, although provide the automatic orthogonal connector routing, produce the routes which may overlap other objects in the diagram.

Popular software packages, Microsoft Visio 2007, and ConceptDraw Pro5, provide the object-avoiding orthogonal connector routing, but in both cases the aesthetic criteria, such as minimizing distance or number of segments, are not taken into account. We cite the conclusion made in the introduction of [18]: "in all current tools that we are aware of, automatic routing of orthogonal connectors uses ad-hoc heuristics that lead to aesthetically unpleasing routes and unpredictable behavior". Agreeing with the latter conclusion as well as with the remarks cited in the Introduction we find the developing of new domain-specific algorithms reasonable.

3. A Modified Shortest Path Algorithm

The most frequently discussed criteria of the aesthetic attractiveness of connectors are length, number of bends, and number of crossings. The first two criteria seem better justified than the last one which seems conditional. Some crossings, e.g. where long edges cross at their middles, is not a negative factor for the perception of the relations indicated by those edges. This objective, however, can be fine tuned to address these corner cases. Having this in mind we start by consider the routing of connectors, focusing on the first two criteria.

Natural candidates for the construction of connectors according to the criterion of connector length are shortest path algorithms. Those algorithms are efficient from the theoretical and practical points of view, and their software implementations were elaborated during intensive, long lasting applications in real world problems. However, the connectors found by a standard shortest path algorithm for the diagrams discussed above frequently are disadvantageous because of many bends. That disadvantage is indeed natural: normally there exist several different paths between two shapes with equal lengths but different number of bends. A trivial solution to find all shortest paths and select one with minimum number of bends is not attractive because of substantial increase of the computation time. Possibly, a new bi-criteria algorithm could be developed for the domain-specific graphs which correspond to the diagrams described above taking into account both criteria – the connector length and the number of bends. However, it would seem not likely to preserve the efficiency of standard shortest path algorithms and software achieved by the many years of refinement. We propose a rather simple modification which enables us to take into account both criteria by using a standard shortest path algorithm.

Let us specify the data for use with a standard version of a shortest path algorithm. The set of vertices of a graph comprises the points denoted in the Figure 1 plus vertices on the shapes marking the ends of the connectors. The set of edges consists of the segments of horizontal and vertical lines between the vertices, if those segments do not cross the shapes. The weights of edges are equal to their lengths. There would usually be several shortest paths in such a graph, and geometrically most of them are of the zig-zag type. This is due to the fact that for such a structure as this graph, the Manhattan distance is used and there are a lot of paths with the same Manhattan distance connecting any two vertices. We also define a modified problem for a graph with the same set of vertices as before but with a different set of edges. The latter also includes the segments of the vertical and horizontal lines with the intermediate points, e.g. in Figure 2 three edges should be considered: (a, b), (b, c), and (a, c). The weight of an edge is equal to the

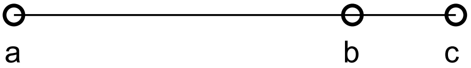


Figure 2. A segment of line represents three edges (a,b), (b,c), and (a,c)

square root of the edge length. In this case, two paths of equal length but comprised of segments of different length can have different weights: the path comprised of small number of long segments will have smaller weight than the paths comprised of large number of short segments. By such a definition of weights, the paths with smaller number of bends implicitly are preferred for being selected by a shortest path algorithm.

The complexity estimates based on asymptotic analysis are not very relevant here since the sizes of problems to be considered cannot be very large otherwise the corresponding diagrams could not be properly surveyed and understood by the user. Nevertheless the complexity estimate is of some interest. Let the size of the orthogonal mesh be $m \times n$, and the number of shapes is denoted by k. The mesh is supposed to be tightly filled: $k = \gamma \times m \times n$, $\gamma < 1$. Assume for simplicity that $m = n$. Then the number of vertices is $O(n^2)$, and the number of edges is $O(n^3)$. Let the shortest path problem in such a graph be solved by Dijkstra's algorithm where the priority queue is implemented as Fibonacci heap. Then the complexity of the algorithm can be estimated as $O(n^3 + n^2 \log(n^2)) = O(k^\alpha),\ \alpha = 3/2$.

The complexity of searching for shortest paths, similar to those searched by the proposed algorithm, can be reduced by taking into account the geometry of the diagram explicitly. However, the gain does not seem counterweighting the loss of flexibility in further modifications of weights on edges taking into account various criteria of the layouts' aesthetics.

4. A Version of the Ant Colony Optimization Algorithm

The modified shortest path algorithm presented in the previous section is primarily oriented to the minimization of the paths' length. In that algorithm, the criterion of bends is taken into account implicitly. The criterion of the number of crossings is not taken into account. The formal involvement of all three criteria by a modification of a known, say the shortest path type, graph algorithm seems difficult. Therefore we start with the general comments on the algorithm selection for a muti-objective optimization problem. The multi-objective algorithms of the type of the classical mathematical programming are efficient for the problems where the objective functions satisfy very restrictive, from the point of view of applications, requirements [12]. Indeed, it seems difficult to find an algorithm of the classical mathematical programming type suitable to the considered problem. As shown in [17], the stochastic algorithms are appropriate for the single objective global optimization of objective functions with various irregularities. The special case of the stochastic algorithms, namely the evolutionary algorithms, are appropriate for solution of various applied multi-objective problems as shown in [5]. Therefore a stochastic metaheuristic algorithm seems appropriate to development of an algorithm for the problem considered. The ant colony optimization (ACO) algorithms are especially oriented to the search for short paths in the complicated graphs [4], [6]. In the recent paper [9] it was shown that the ACO algorithms are efficient in solving the bi-criteria traveling sales person problem. Following the arguments above we have developed a ACO specified for the considered problem. The version of the proposed ant colony optimization algorithm differs from the standard version in the amount of pheromone placed on the path traveled by the ant: it is inverse proportional to the path length, number of bends, and number of crossings squared in contrast to only taking into account the path length. Below we present a description of the algorithm.

1. Mark each edge in the graph with a pheromone value for that edge. In the beginning this value is one. In this way, at the start, all paths are equally likely to be chosen. It is possible to set pheromone values to something other than one at the start. For example we could use some other algorithm, or even ACO itself, maybe with different parameters, to generate a set of paths, and use those paths to modify pheromone values in advance, thus giving the algorithm a head start. Then, the ant colony optimization algorithm could be used to fine tune these paths by, say, emphasizing reduction in the number of bends or crossings.
2. Generate 10 random paths using the procedure outlined below.
   (a) In the beginning the path consists of the starting vertex only.
   (b) Generate successors for the last vertex in path.
   (c) Assign probabilities to each successor by taking pheromone values for edges going from the current vertex to the successor. Normalize these probabilities to add to one.
   (d) Select one of the successors at random according to the probability distribution generated in step (c).
   (e) Attach the selected successor to path.
   (f) If the new vertex is the final one then terminate.
   (g) Go to step (b).
3. For each path run the procedures outlined below.
   (a) For each edge in path add $1/(length + folds + intersections^2)$ to pheromone value of that edge. This is similar to the way that fitness function is computed in the genetic algorithm.
4. For each edge in the graph multiply the pheromone value by 0.9, which simulates pheromone evaporation. In this way the paths that are the shortest, have the least bends and least intersections with other paths will tend to be used most since they will get the highest pheromone values. The other paths will have their pheromone values constantly reduced until it reaches levels so low that almost no ant will choose them.
5. Repeat from step 2 a specified number of times. In the experiments the number of times was set to 500. However values of 200 or even lower were found to be sufficient for the paths to settle.

5. Computational Experiment

The proposed algorithms are implemented in C++, and their performance was evaluated experimentally. Both algorithms are sufficiently fast in the sense that they produce results

Table 1. Mean values and standard deviations of the criteria of solutions found by the modified shortest path algorithm

k	$L(means)$	$L(std)$	$N_b(means)$	$N_b(std)$	$N_c(means)$	$N_c(std)$
6	6.2200	1.9467	6.4800	2.4141	2.0000	1.8257
8	8.1100	2.2514	9.8000	2.7340	3.1000	2.0865
10	10.1500	2.3067	12.8600	3.0385	4.5200	2.4923
12	12.4500	2.8652	16.9400	3.6785	6.8200	3.1120

Table 2. Mean values and standard deviations of the criteria of solutions found by the ants colony optimization algorithm (first mode)

k	$L(means)$	$L(std)$	$N_b(means)$	$N_b(std)$	$N_c(means)$	$N_c(std)$
6	6.8239	0.6681	7.2002	0.7881	1.2925	1.0163
8	8.9890	0.7383	10.3854	0.7362	2.4895	1.4287
10	11.4066	0.9201	13.4010	0.7897	3.9490	1.9956
12	14.2724	1.0055	17.4508	0.7265	6.8063	2.5723

Table 3. Mean values and standard deviations of the criteria of solutions found by the ants colony optimization algorithm (second mode)

k	$L(means)$	$L(std)$	$N_b(means)$	$N_b(std)$	$N_c(means)$	$N_c(std)$
6	7.6706	2.9545	8.9709	3.0115	1.1089	1.2598
8	9.9101	3.3273	12.106	3.3389	2.6572	2.4788
10	12.4737	3.5373	15.1579	3.7099	4.3245	3.7035
12	15.4569	4.3247	19.1217	4.4041	7.7382	5.2840

in time not noticeable by the user. Statistics for the quality criteria of connectors were collected after solving randomly generated test problems.

The modified shortest path algorithm has been applied for the reduced graphs containing only edges at the center of alleys; as can be seen from Figure 1 alleys consist of either three paths or of two paths in our particular problem. It is supposed that the coincident parts of connectors can be separated at the final step of the connectors' refinement, e.g. by the procedure of "nudging" [18].

ACO algorithm has been used in two modes. The first mode corresponds to the reduced graph described in previous paragraph. The second mode corresponds to the original graph.

Locations of shapes were generated at the nodes of the rectangular mesh randomly with uniform distribution. Randomly selected pairs of shapes were connected. The size of the mesh was 3×5. The number of shapes was 6, 8, 10 and 12, representing problems of increasing complexity. While this may seem artificial in the context of BPMN diagram drawing, however we are interested in more abstract properties of the algorithms. The following parameters of the found connectors were evaluated: the total length of connectors (L), the number of bends (N_b), and the number of crossings (N_c). The mean values ($means$) and standard deviations (std) of these parameters were computed using the data of 100 solved problems which were generated randomly as described above. In the case of ACO algorithm, the experiment was repeated 100 times for each of 100 sets of nodes. Means and deviations were then averaged. The results are presented in the Tables.

The experimental results show that both algorithms are of similar efficiency. However the ACO algorithm is more flexible with respect to the increase of number of the criteria considered. Further investigation is supposed including the results of a psychological experiment aimed at the quantitative assessment of the importance of the potentially applicable criteria [21].

Conclusions

Two algorithms of routing of orthogonal connectors, supposed for the aesthetically pleased visualization of edges of special graphs, are proposed. Both of them are sufficiently fast to be useful in the visualization of graphs related to the modeling of the business processes of SME's. The aesthetic criteria of the found connectors are evaluated quantitatively. The experimental results show that the ant colony optimization algorithms are promising for the solution of the considered multi-objective graph optimization problem.

Acknowledgements

The support by Agency for Science, Innovation and Technology (MITA) trough the grant Nr.31V-145 is acknowledged.

References

[1] Battista, G. D., Eades, P., Tamassia, R., Tollis I.G.: *Graph Drawing: Algorithms for the Visualization of Graphs*, Prentice Hall (1999).

[2] Bennett, Ch., Ryall, J., Spalteholz, L., Gooch, A.: The Aesthetics of Graph Visualization. In: Cunningham, D. W., Meyer, G., Neumann, L.(eds.) *Computational Aesthetics in Graphics, Visualization, and Imaging* (2007), 1–8.

[3] Chen, H.-Y., and Chang, Y.-W.: Global and Detailed Routing. In: L.-T. Wang, Y.-W. Chang, and K.-T. Cheng, (eds.) *Electronic Design Automation: Synthesis, Verification, and Testing* (ISBN: 0123743648), Elsevier/Morgan Kaufmann (2008).

[4] Colorni, A., Dorigo, M., Maniezzo, V.: *Distributed Optimization by Ant Colonies, actes de la premiere conference europeenne sur la vie artificielle*, Elsevier Publishing, Paris, France (1991), 134–142.

[5] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons (2009).

[6] Dorigo, M., Birattari, M., Stutzle, T.: Ant Colony Optimization, Technical Report No. TR/IRIDIA/2006-023 (2006).

[7] Effinger, Ph., Jogsch, N., Seiz, S.: *On a Study of Layout Aesthetics for Business Process Models Using BPMN*, Lecture Notes in Business Information Processing, Vol. 67 (2010), 31–45.

[8] Effinger, Ph., Kaufmann, M., Siebenhaller, M.: *Enhancing Visualizations of Business Processes*, Lecture Notes in Computer Science, Vol. 5417 (2009), 437–438.

[9] C. Garcia-Martinez, O. Cordon, F. Herrera. A taxonomy and an empirical analysis of multiple objective ant colony optimization algorithms for the bi-criteria TSP. *European Journal of Operations Research* **180** (2007), 116–148.

[10] Kitzmann, I., König, Ch., Lübke, D., Singer, L.: A Simple Algorithm for Automatic Layout of BPMN Processes, In: *1st International Workshop on BPMN (CEC-09 – 11th IEEE Conference on Commerce and Enterprise Computing)* (2009), 391–398.

[11] Lee D., Yang C., Wong C.: Rectilinear paths among rectilinear obstacles. *Discrete Applied Mathematics* **70**(3) (1996), 185–216.

[12] K. Miettinen. *Nonlinear Multiobjective Optimization*, Kluwer Academic Publishers (1999).

[13] ORGSOFT, http://www.orgsoft.lt.

[14] Owen M., R.Jog.: BPMN and Business Process Management (2003), 1–27, http://www.bpmn.org.

[15] Purchase, H. C., McGill, M., Colpoys, L., Carrington, D.: Graph drawing aesthetics and the comprehension of UML class diagrams: an empirical study. In: *Australian Symposium on Information Visualization, Conferences in Research and Practice in Information Technology*, Vol. 9 (2001).

[16] Purchase, H. C.: Metrics for graph drawing aesthetics. *Journal of Visual Languages and Computing* **13**(5) (2002), 501–516.

[17] A. Törn, A. Žilinskas. *Global optimization*, Lecture Notes in Computer Science, Vol. 350 (1989), 1–255.

[18] Wybrow M., Marriott K., Stuckey P.: *Orthogonal Connector Routing*. Lecture Notes in Computer Science, Vol. 5849 (2010), 219–231.

[19] Yang C-D., Lee D., Wong C.: Rectilinear Path Problems among Rectilinear Obstacles Revisited, *SIAM J. Comput.* **24** (1995), 457–472.

[20] Žilinskas A., Žilinskas J.: Optimization based visualization, In: C.Floudas and P.Pardalos (Eds.), *Encyclopedia of Optimization*, Springer (2009), 2785–2791.

[21] Žilinskas A., Mackutė-Varoneckienė, A., Varoneckas A.: Weighing Criteria of Aesthetic Attractiveness of Layouts of Business Process Diagrams, In: *Proceedings of STOPROG 2012*, in print.

Integration of Business Modeling and IT Modeling

Girts KARNITIS[a,1], Janis BICEVSKIS[a] and Jana CERINA-BERZINA[b]
[a] *University of Latvia, Raina 19, LV-1586 Riga, Latvia*
Email: Girts.Karnitis@lu.lv, Janis.Bicevskis@lu.lv,
[b] *Datorikas Instituts DIVI, A. Kalnina str. 2-7, LV-1050 Riga, Latvia*
Email: jana.cerina_berzina@di.lv

Abstract. The paper covers our research in improving business and IT process modeling. Practical experience shows that universal modeling languages such as UML and BPMN are IT oriented and are difficult to accept by business people. Business models need to use such hard to formalize business terms as legal acts, informal activities descriptions and comments etc. We propose novel approach how to create or choose modeling tools according to model usage. We propose 4 usages for modeling: model to understand business processes, model for business process definition, model as information system specification and model that can be interpreted by computer. All these four models can be perceived as four highly interconnected parts of one single model. We propose to create unique domain specific language for each model. We analyze functions specific for each model type. Authors last six years practical experience for creating and practical introducing related business and workflow oriented information system models are analyzed.

Keywords. Business process modeling, domain specific languages, modeling

Introduction

Business process modeling and information system specification have number of problems irrespective of wide research in a field of Model Driven Software Development (MDSD) [1] and Model Driven Architecture (MDA) [2]:

- Universal modeling languages UML and BPMN [3, 4] do not provide language easily understandable by both – IT specialists and business people. IT specialists propose to use UML than can describe operation of IS objects. In such case process part outside the computer stays weakly specified. Therefore business people cannot adequately describe their requirements.
- Model syntax prevails over model's meaning – semantic. Universal modeling languages concentrate on graphical symbols usage laws and do not explain symbols semantic meaning. It's even impossible, because in each particular case semantic may differ. As a result process semantic is hard to put into the models and to read it and models allow different interpretations.

[1] Corresponding Author. As. professor, University of Latvia, Raina 19, LV-1586 Riga, Latvia; Email: Girts.Karnitis@lu.lv.

As a result software requirements specifications and job descriptions are created in form of a written document containing illustrative graphs and schemes.

Our novelty approach is to concentrate to model semantic according to specific model usage. We have identified 4 important model usages – informal business description, precise business model that can be used as work instructions for people, business process model that serves as requirement specification for IS development, business process model that is used by IS to operate. In details – we propose to add semantic to each symbol used in diagrams according its usage. Widely known modeling languages, such as UML activity diagrams or BPMN can be adopted to acquire abovementioned goals. But because of specific requirements (link with information in existing IS, consistency checks) we decided to use specific DSL.

Universal modeling languages specify only syntactic structure of diagram and do not specify semantic meaning. We propose assign to activities/tasks not only unique names but also attach informal descriptions defining semantic meaning of each activity/task. In this case diagrams syntax can conform to UML Activity Diagrams or BPMN or domain specific language (DSL), but semantically they are supplemented with these informal description and shows essence of modeled system. It brings us to specific DSL [5] that allows us to understand models in context of specific system diagrams made with universal modeling tool.

This technique is used in number of projects in Latvia. Result shows that proposed approach allows minimize problems with connections between business people and IT specialists. In this paper we explain proposed approach with small process and 4 usage scenarios and 4 different types of models of this process.

The structure of the paper is as follows. Chapter 1 contains four usage scenarios and four different DSLs for these usage scenarios. Chapter 2 contains analysis of properties of proposed DSLs. Chapter 3 contains our experience with usage of these DSLs.

1. Types of Models

During analysis and modeling number of organizations we have developed 4 different model usage scenarios. In this chapter we analyze these model usages and show how to add semantics to the models. As a result we get 4 DSLs that allows to model wide spectrum of workflow oriented systems:

1. Business process description – due to graphical illustration of the processes the model contributes to better understanding of the nature of business processes [6].
2. Business model determines definition of a process – formalized language determines process steps, implementation and reports, communicated among process executors, thus creating precise descriptions of functions carried out by the employees. Moreover, process implementation steps are closely linked to the organizational structure – process executors and their activities under respective normative acts. Clearly, IT specialists proposed UML only partly meets the demands of above-mentioned non-formalized descriptions.
3. Business model serves as requirements specification for the IS development – description of business process steps liable to IS support are included in the model as well as all process implementation rules and algorithms.
4. Business process model creates the base of IS operations' description [7]. Such situation is common in recordkeeping and event-oriented systems, where work-

flow is described by graphical diagrams in specialized Business Process Modeling Language (BPML), but software is developed by generating Business Process Execution Language (BPEL) commands from the model in BPML. Thus, we consider the work-flow description language BPML to be domain specific language.

1.1. Model as the Process Description

Usually, the initial organizational processes are missing description or descriptions are anchored in legislative and normative acts and instructions. Frequently, these instructions are incomplete or even contradictory and incomprehensible. Process models are created to improve this situation. First step is to create a model that describes process and serves as a descriptive material for organizational process. Users of these models are usually top business managers valuing mainly comprehensibility and simplicity of the model. Here the common principle requires intuitive guess of model's syntax or a minimum training. Main requirement of such models – users without deepest understanding of modeling language must intuitively understand process described in the model. Main components of such model are graphical diagrams illustrating business process activities and their implementation sequence, as well as informal textual descriptions of activities attached to diagrams. Usually, the informal descriptions include references to normative acts, document templates and similar documents. Graphical diagrams and informal descriptions include also other related information such as information about process executor, deadlines and necessary technical resources.

As an example for the rest or the paper we will use registration of enterprises to allow them to perform specific kind of entrepreneurship. A law in Latvia as in many countries stipulates that to carry out some specific kind of entrepreneurship, for example to provide communication services, special license must be received. Figure 1 shows the highest level model for such process. Enterprise must address application to responsible state institution. In this case responsible institution is Public Utilities Commission (PUC). PUC considers application in 15 days and issues license and registers enterprise or rejects application in a case if some requirements are not fulfilled.

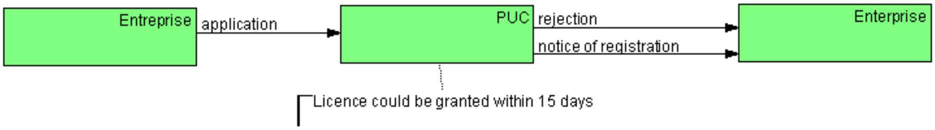


Figure 1. Enterprise registration process model

This simple picture can be formatted according to syntax of different modeling language. Still the meaning of the process can be understood by reading the description in precious paragraph. It means that main goal of modeling language is to bind process semantics with graphical picture. Frequently semantic is described via informal descriptions and normative acts that is not the part of the model itself. Rather process description and normative acts are stored into separate files, different web sites or paper documents. It is hard to use models without attaching these descriptions.

Figure 2 pictures the same process from PUC viewpoint. In this case registration process is divided into 4 consecutive steps – registration of application, enterprise information check, decision making weather to issue license or not, and license registration in register in case license is issued. This model is not in great interest for enterprises, but can be valuable for PUC employees to carry out their jobs. Similarly to previous example, this model is hard to understand without additional information (semantics explanations) how to carry out each step.

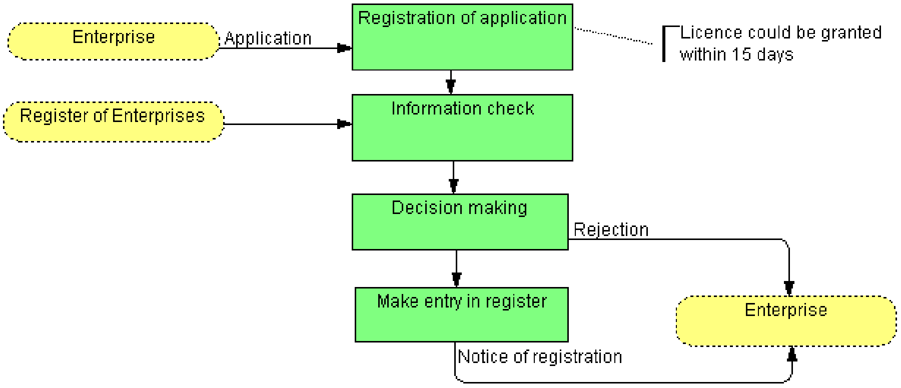


Figure 2. Enterprise registration process model from PUC viewpoint

The most important - theses models are easily comprehensible and well presented, since their main objective is to provide good insight and receive management's approval for publication. However, there is also a main deficiency – these models are rather too abstract and fail to be completely consistent. These models usually have low formalization level. Actually, any graphical tool, such as Microsoft Visio, that provides unlimited choice of symbols, layouts and visualization possibilities, would be suitable for drawing a model of such sophistication level.

Representation of organizational processes by informal schemes, operational descriptions and diagrams is a common modeling approach anchored in process penetration and caused by requirements ISO quality management standards. These models, which are developed for solely the purposes of quality management, contain a certain degree of inaccuracy in order to use them for development of information system operations or monitoring enterprise performance indicators. As a result, the ultimate goal of quality management cannot be met, since there is a lack of objective performance assessment and informal schemes of processes can be used only by limited number of users.

IT specialists can use such model for better understanding of processes and industrial specifics during IS development initial phase. Informal nature of the model constraints its use for IS development and also cannot be used as precise official instructions.

1.2. Model for the Definition of Process

The next step of process modeling development would be to develop business process models that define business processes and provide a precise instruction for process execution [1]. This approach sets significantly higher and sophisticated requirements for model's formalization, precision and consistency than in the case of models developed solely for the purpose of process understanding. These models should have a considerable degree of precision for employees to be able to interpret the model information precisely and execute the process steps correctly. However, the level of precision and formalization sufficient for employee's interpretation might not be sufficient for the implementation of such model in information system or for the automated execution of the model by the information system.

In our example, the executable model contains additional information about the process steps, activities carried out in each, including screenshots of IS forms and reports where information must be entered and can be read during this step. Technically this information can be seen with the corresponding model editor tool that would allow attract the appropriate screen forms or documents to each process step. Criteria for the models in this group - the user has access to all information necessary to process each step in real life.

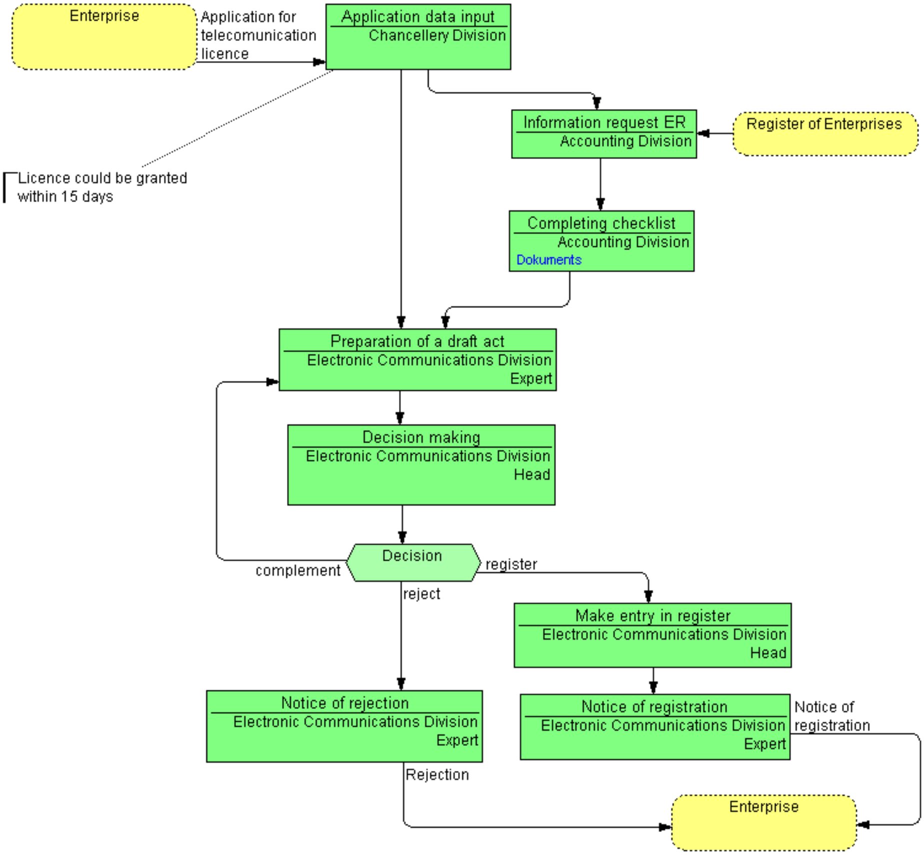
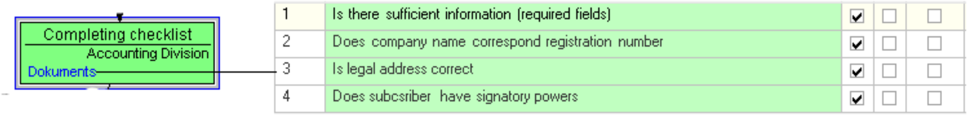


Figure 3. Executable model of enterprise registration

Models that are used to define processes contain graphical diagrams and informal descriptions as well as references (links) to external documents (precise references to paragraphs in normative acts and document templates), references to functions of external information systems (screen forms) and user instruction on filling these forms

(Figure 4), references to objects in external information system (e.g., process steps executor in organizational structure). Model for process definition cannot bear any inconsistency, therefore consistency verification is crucial. Conditions of consistency check-ups should be defined in the same modeling language and check-up should be performed by the modeling tool. It should be emphasized that consistency conditions are domain specific, which means – theses conditions are specific to organizational and business requirements of a particular organization.



1	Is there sufficient information (required fields)	☑	□	□
2	Does company name correspond registration number	☑	□	□
3	Is legal address correct	☑	□	□
4	Does subcsriber have signatory powers	☑	□	□

Figure 4. Activity with external link – checklist form

Main users of model are employees, which use the model as an instruction for everyday execution of business processes. Above mentioned model does not contain precise data structures, therefore it cannot be automatically executed and only partially contains information necessary for IS requirements specification.

1.3. Model as IS Requirements Specification

Model described in the previous sections were used to define business people view of processes taking place in organization. In abovementioned model business processes are described maximally precise from business point of view so that employees can accomplish them. Business people and also IT people are able to understand language used in such models. However, these definitions and descriptions are not sufficient in order to describe operations of information system – these models must be complemented with data structures and other information required to develop IS. Such models we will call IS diagrams.

Usually IS diagrams are created by the information system analysts. IS diagrams contain descriptions of software screen forms, reports and automated tasks – algorithms. IS diagrams describe conceptual data level, which can be detailed further to the level of physical data model. IS diagrams continue detailing the business process steps from the point where business process diagrams have reached their constraints. For example we take process step "Make entry in register" form previous model in Figure 3 and detail it. We specify this step as two steps and add IS menu items needed to perform this step (Figure 5).

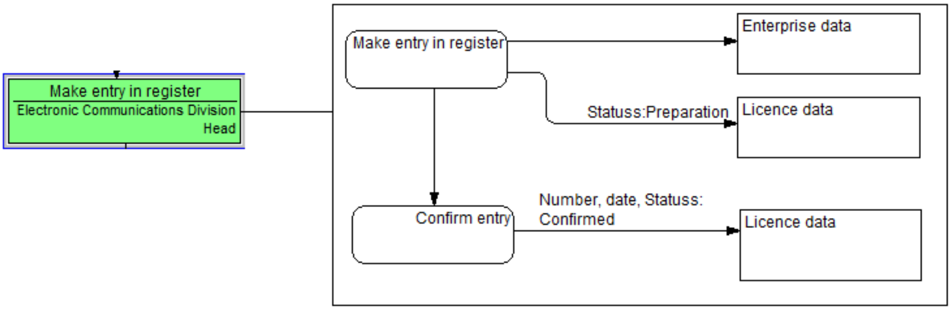


Figure 5. Refinement of "Make entry in register" process step

IS diagrams have several ways of applicability:

- IS diagrams serve as a basis for development of software functional requirements containing data input, data output, data processing and data verification.
- IS diagrams determine, which business process fragments should be implemented in information system and became fully automated and which should remain manual even after information system is introduced. For example, if a business process foresees verification of a signature on submitted document then the employee will perform the verification manually. IS can only verify input values and information consistency.
- IS diagrams allows to make data usability analysis. For, example, if we would like to change processing logic of data object we can create view where we can see all IS activities that use or influence these data. A different view can be created in order to monitor status change of an object in order to indicate activities related to object status and its changes.
- IS diagrams can serve as a basis for creation of test cases for software testing. Use of both, business process model and IS diagrams, can easily solve the problem of software testing completeness - business process model can be used as testing model.

In summary, if business requirements are gradually and successfully developed into IS requirements, organization obtains a very precise description of organization's information system and IT specialists obtain functional requirements for the system implementation. By using and maintaining business process and IS models in a long term organization obtains significant information on impacts and costs of changing or adding new business functions to organization.

1.4. Model as IS Component

Current trends in information system technological development can be characterized by less technical coding, more precise specification resulting in software development "without programming". Modeling is the integral part of this process.

IT trend, where modeling builds the foundation to systems operations, is represented in workflow-oriented systems. A workflow indicating process steps - activities is developed according to organization's business process. In a workflow significant role is played by executor – user, which has an assigned role in organizational structure. Process steps perform data processing of involved object. Work-flow provides possibility to store data throughout the object's life-cycle. Such concepts and terms as task deadlines, actual status of object, possibility to create documents automatically during workflow execution are very significant in work-flow oriented systems.

There are two specific tasks for models, first, it should be rather simple and comprehensible for employee to understand workflow, second, it should be sophisticated enough in order to contain all information significant for operation of information system. In principal, these are two different views of the same workflow diagram.

Which would be the most appropriate language for creation of workflow description? One possibility is use UML or BPMN. Unfortunately, in the UML and BPMN models because of their universality is difficult to incorporate domain semantics. In addition, the UML and BPMN are overloaded with many technical features, which are needed for the development of IS in specific cases, but often are redundant for a particular system and is difficult to understand by the business people. Frequently acceptable option is to use a DSL, which allows you to build models for both – business needs and for the IS implementation.

Authors of this article have developed domain specific language BILINGVA [7], with the purpose to describe workflows. Diagrams illustrate object statuses, process steps (events) and sequences of process steps. The language provides possibility to describe also decisions – branching points in processes, define deadlines and execute processes parallel. After a model is created the information stored in diagrams is transported to the information system repository and system's configuration tables. By operating information system its engine interprets information obtained from diagrams automatically, thus ensuring system functionality defined in the model.

Figure 5 contains executable enterprise registration process. Unlike diagram from previous chapter, all activities in this model can be executed by IS. The model does not contain objects that are not covered by IS such as "Enterprise" and "Enterprises register". The model also does not contain manual operation such as document review. In BILNGVA syntax arrows are activity/task but boxes are object states. Parallel execution also can be shown. After application, registration, enforcement activity has been transferred to the Accounting Division, operating as sub process in parallel with the main process, and shall be executed by an expert of Electronic Division.

Time limit of the process is represented as two events – "begin time control" and "test time control". In event "begin time control" event "Time to register – 15 days" is stored. IS form shown to user in a case time control rule is infringed is implemented into model interpreter and has not specified in the model. This helps to create user-friendly interface for execution of concrete activities. Such user friendliness is hard to achieve in a case if universal solutions are used.

Diagram in Figure 6 is a user-oriented view, where the notation proposed in [7] is used. To ensure correct work of system process modeler completes attribute set for each process step that is necessary for correct operation of system. For example each step is characterized by the following parameters – does step can make document, if so, than document template must be specified, does some document can be uploaded to system during the step, does specific checklist is fulfilled during the step, step executors roles – manager, responsible employee or anyone. Does responsible employee can be set in this step? In which conditions step can be executed? Step status attribute set describing which departments see this status, status number - to sort statuses in user interface in logical order, status group. Time limit is additionally characterized by following attributes: is time limit in calendar days or working days, is time limit automatically or manually counted, how many days before deadline system issues warning about deadline, which departments are bound to the deadline.

Returning to applicability aspect, a detailed workflow diagram provides framework for system's operations, a special view serves as a user's manual, which can be complemented by explanatory documents referenced to certain steps.

The future perspective includes task – to implement graphical representation of process step execution in business process diagram visible to system user. For example, by providing simplified process model on organization's homepage the customers could track and follow execution status of provided service.

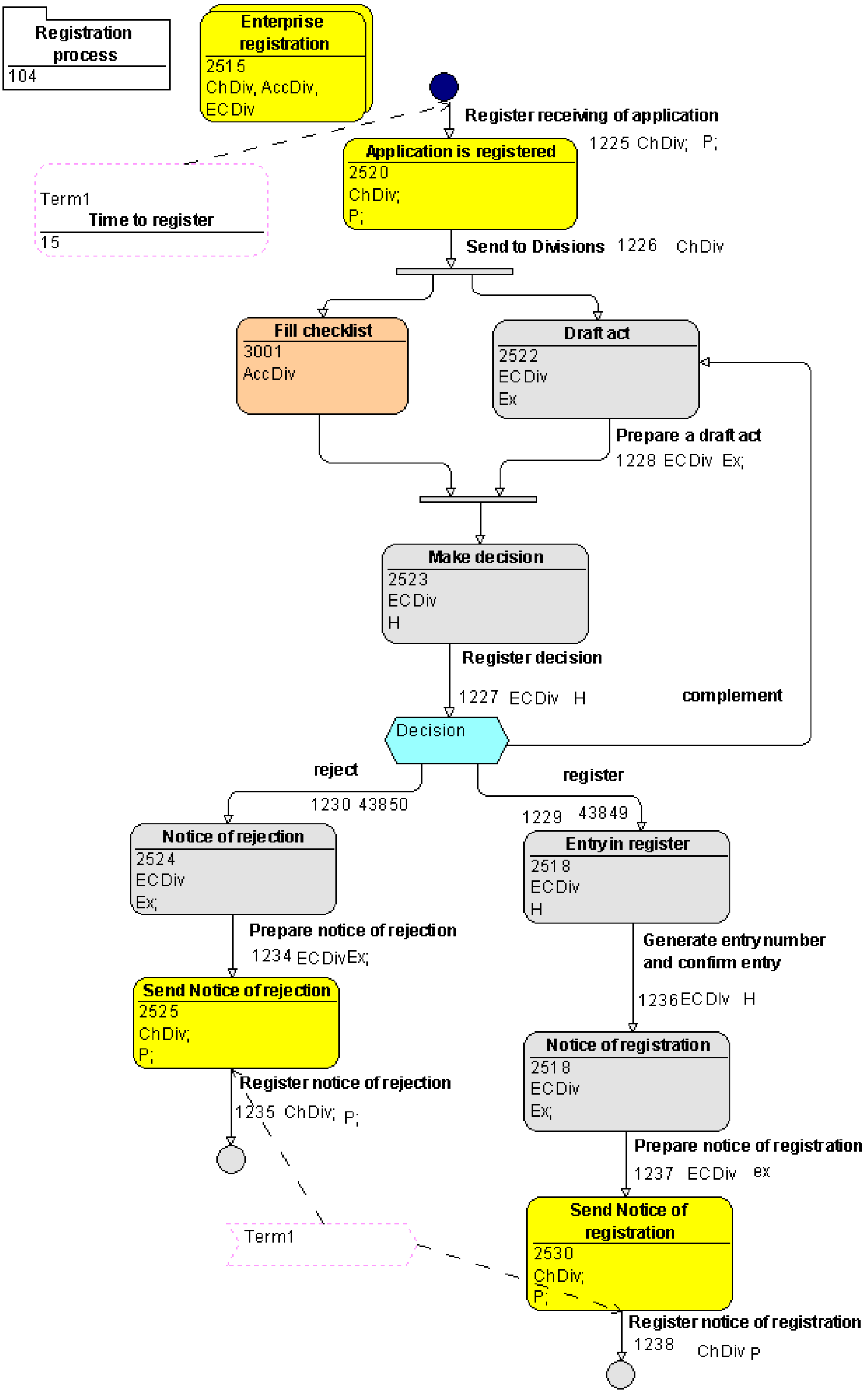


Figure 6. Executable process model

2. Components of Model

2.1. Models Attributes

Each of four models described above have their own applications and users. Consequently, these models differ in the semantics of elements and attributes of the elements. This fact is presented in Table 1 where some attributes of the process steps of various models are given.

Table 1. Attributes of process steps

	Purpose of process	Definition of process	IS requirements specification	IS component
Semantic precision	Very low formalization	high formalization	very high formalization	100% formalization
Syntax precision	Very low formalization	high formalization	very high formalization	100% formalization
Step name	Yes	Yes	Yes	Yes
Step description	Might be	Yes	Yes	Only for human. IS does not need.
Performer	Might be	Yes	Yes	Yes
Time limits	Might be as informal comments	Yes	Yes	Yes, in fully formal form

Documents bound to the model can differ according to the model type:

- Model for purpose of process – process regulating documents are attached as external files. Usually documents are attached to the whole process not to specific process step.
- Model for definition of process –regulatory documents preferable attach to specific process step. IS screenshots, document templates, user manuals and job descriptions are attached to the process steps.
- Model for IS requirements specification –specific chapter regulating process step preferable attach to process step. IS forms that must be fulfilled in this step, description of used data (data model), consistency checks, error messages and other information necessary for requirements specification are attached to the process step.
- Model as IS component – document templates used in this step to generate documents are attached to the process step.

Other attributes can be added to the model according to the domain specific.

2.2. Model Views

Model describes the operation of a single object - the organization, its processes and their implementation. Our approach offers to build a single model with common information and add different views to model for different users and different applications. Various pieces of information from the model (different views of the model) are necessary for business representatives, customers, quality service, IT professionals, and so on. The above mentioned applications show how wide are usage of models. We will mention number of other applications in this chapter, which requires their own solutions.

One way to use model is different views for different applications. Example of view is series of activities performed to accomplish specific service (for example, registration of enterprises to allow them to perform specific kind of entrepreneurship). Another example of a view - the process steps that specific position is interested in - for example, Manager view of the processes, which include not only the steps Manager performs, but also actions that are performed before and after him. Another example of view - the quality system view of the organization's operations, which includes the entire range of activities, but rather superficial, omitting a number of components.

We may say that Manager is interested in some of the process execution to the smallest detail, but in some other process execution is actually not interested at all. By contrast, the quality department is interested in all the processes, but to the highest top level without going into detail.

It is unacceptable that implementation of the above mentioned views would be generated in a separate model, unrelated to the common organizational model. In order to maintain consistency each such view should be implemented as a subset of the overall model. Currently, such views can be created manually and consistency checks can be built into tool. At present, an unresolved problem is the view definition and automated creation from the integrated model.

Moreover, the model can be applied in a further way – service diagrams can be developed based on functionality model. Service diagrams illustrate certain process steps to be carried out in order to provide a particular service to customers. For example, a help-desk inspector at one stop agency provides services to a number customer with various demands. It is impossible for inspector to precisely know by heart every process step of a number of services applicable for each customer case. Therefore the inspector uses process model that includes precise instruction for each service case.

A different reporting possibility is a further way of model applicability, for example, reference of process steps to certain legislative acts. In case of amendments of legislative acts these references easily indicate process steps that should be changed. UML and partly also BPMN can be used as a language to meet the requirements of such models for these languages provide precise definition for process execution. However, missing possibility of adding domain specific information to model makes use of theses languages problematic, for example, would it be possible to easily added references to paragraphs of normative acts if a process step is defined in UML or BPMN? Moreover, since the consistencies usually are domain specific, the use of universal modeling languages makes keeping of model consistencies problematic. The elimination of this problem would require development of specific tools in order to ensure consistency in model created by universal languages.

3. Real Life Experience

Authors has several years of experience in the use of DSL in seven organizations in all four analyzed above applications. Usually each case consisted of three steps. The first step was DSL definition, typically through business process modeling together with the company's business professionals. Usually new DSL was based on some already existing DSL and a new definition seemed to add extra attributes and attach domain-specific input / output documents. DSL definition virtually at the same time meant the definition of modeling tool. That was made possible by using a DSL tool building platform [8, 9, 10]. In the third step, the business model was developed using newly created DSL and modeling tool. This model was then transferred to a business-specific application. Accumulated experience has shown this approach strength:

- Each new DSL definition reveals new requirements for organizational model, such as model's elements linkage to the sector-specific document flow. Identified requirements can be incorporated into DSL and model.
- General purpose modeling language cannot describe all the sector specific requirements, for example, it's hard to access to non-formal process descriptions stored in the enterprise IS. Similarly, if universal modeling language is used then it is hard to link document samples available in company IS or website to the process descriptions.
- It was especially surprising to discover very positive attitude of users towards graphical specifications which users preferred to use as information system exploitation manual for it contained more precise information in a more comprehensible way, thus replacing old-fashioned user manuals.

Modeling technology is a new type of activity for business people. It takes time and the gradual acquisition of modeling techniques. At first, business people are able to create informal models (informal process description models) and only after some time they are able to move to more formal models (formal models of process definition). In our case adoption of modeling technique and models in different organizations took 2-4 years.

4. Conclusions and Proposals

The experience of DSL's use allows make such main conclusions. Strength of DSL:

- DSL allows create single unified model within organization for all four different usage scenarios mentioned in this paper and allows different views to the model.
- Users can easily understand meaning of models and use models if business process semantic is bind to model objects. Syntax specific usually does no matter so much.
- DSL building is one of easiest way to bind semantic of specific domain to model.
- DSL definition and modeling tool definition platform plays one of mayor role for DSL usage. It is practically impossible to implement in real life methodology we describe without such platform. Platform we use [8,9,10] ensures DSL definition and modeling tool definition in relatively small time frame (approximately 3 month).
- Survey shows that end users definitely prefer graphical diagrams instead of traditional text documents. Experience shows that one graphical diagram can contain the same amount of information as 10 pages of textual instructions.
- Proposed process modeling methodology is examined in workflow type systems. For other types of systems the situation may differ.

Recognizing the benefits we also must admit problems arising from the usage of DSL:

- Definition of DSL and development of modeling tools requires involving high qualification specialists.
- Enterprise specific DSL development and business process definition is individual as enterprise specific IS development. Clients must support not broadly recognized DSL technologies, in order to develop and maintain DSL, modeling tool and company models.
- DSL defined for needs of one company is hard or impossible use in other company even if the companies' profiles are very similar. Experience so far demonstrates, that DSL for each enterprise contains nuances specific for each enterprise. Previously developed DSLs can be used in the very beginning of modeling and help to recognize specific of new domain.

We agree to the supporters of MDSD about growing role of models in IS development process [11] and we plan to study modeling and usage of DSL as one of method for company business process arranging and IS developing more closely in near future.

Acknowledgement

The work is supported by a European Social Fund Project No. 2009/0216/1DP/1.1.1.2.0/09 /APIA/VIAA/044.

References

[1] T. Stahl and M. Markus, *Model-Driven Software Development*, John Wiley & Sons, 2006.

[2] *MDA Guide Version 1.0.1*. OMG. Available from: http://www.omg.org/docs/omg/03-06-01.pdf.

[3] *BPMN – Business Process Model and Notation*. OMG. Available from: http://www.bpmn.org.

[4] *UML - Unified Modeling Language*. OMG. Available from: http://www.uml.org.

[5] J. Bicevskis, J. Cerina-Berzina, G. Karnitis, L. Lace, I. Medvedis, and S. Nesterovs, Practitioners view on domain specific business process modeling. In: J. Barzdins, M. Kirikova, editors, *Databases and Information Systems VI,* Selected papers from 9th International Baltic Conference DB&IS 2010, IOS Press, 2011, 169-182.

[6] A. Schleicher, High-level modelling of development processes. In: B. Scholz-Reiter, H. D. Stahlmann, A. Nethe, editors, *Process Modelling*. Springer, 1999, 57–73.

[7] J. Cerina-Berzina, J. Bicevskis, and G. Karnitis, Information systems development based on visual domain specific language BiLingva. In: T. Szmuc, M. Szpyrka, J. Zendulka, editors, *Advances in Software Engineering Techniques*, Lecture Notes in Computer Science **7054** (2011), Springer, Berlin, 124-135.

[8] J. Barzdins, A. Zarins, K. Cerans, A. Kalnins, E. Rencis, L. Lace, R. Liepins, and A. Sprogis, GrTP: transformation based graphical tool building platform. In: *Proceedings of the MoDELS 2007 Workshop on Model Driven Development of Advanced User Interfaces*, Nashville, Tennessee, USA, October 1, 2007. Available from: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-297/.

[9] J. Barzdins, S. Kozlovics, and E. Rencis, The transformation-driven architecture. In: *Proceedings of DSM'08 Workshop of OOPSLA 2008*, Nashville, USA, 2008, 60-63.

[10] J. Barzdins, A. Zarins, K. Cerans, M. Grasmanis, A. Kalnins, E. Rencis, L. Lace, R. Liepins, A. Sprogis, and A. Zarins, Domain specific languages for business process managment: a case study. In: *Proceedings of DSM'09 Workshop of OOPSLA 2009*, Orlando, USA, 2009. Available from: http://www.dsmforum.org/events/DSM09/.

[11] L. Cao, B. Ramesh, and M. Rossi, Are domain-specific models easier to maintain than UML models? *IEEE Software* **26**(4) (2009), 19-21.

Using Functional Characteristics to Analyze State Changes of Objects

Uldis DONINS, Janis OSIS[1], Erika ASNINA and Asnate JANSONE
Department of Applied Computer Science, Riga Technical University, Latvia

Abstract. Event-driven software systems continuously wait for occurrence of some external or internal events. When such event is received and recognized, the system reacts by performing corresponding computations which may include generation of events that trigger computation in other components. After the event handling operation is complete the system returns to the waiting state for the next event occurrence. The response to the received event depends on the current state of the system and underlying objects and can include a change of state leading to a state transition. The state changes and transitions within a system can be formally analyzed by using functional characteristics of Topological Functioning Model (TFM). TFM captures system functioning specification in the form of topological space consisting of functional features and cause-and-effect (i.e. topological) relations among them and is represented in a form of directed graph. The functional features together with topological relationships contain the necessary information to create State diagram which reflects the state changes within system.

Keywords. Topological functioning modeling, functional characteristics, objects, states

Introduction

The behavior of an object over time could be surmised by analyzing system use-case descriptions, activity diagrams, or other software design artifact. To avoid surmising the state change of objects in system, a State diagram is used [1, 2]. The state diagram is part of the Unified Modeling Language (UML) [3]. The application of design models provide better understanding of proposed solution and allows making better decisions concerning the implementation details. Additionally, the model driven development has been put forward to enable development, validation and transformation of syntactically and semantically complete models, thus allowing source code generation automation. In such way models are promoted as the core and main artifact of software design and development.

Despite the presence of UML and a number of software development methods, the way the software is built still remains surprisingly primitive (by meaning that major software applications are cancelled, overrun their budgets and schedules, and often have hazardously bad quality levels when released) [4]. This is due that the very beginning of software development lifecycle is too fuzzy and lacking a good structure since the software developers has limited analysis and modeling of systems [5]. Instead

[1] Corresponding Author: Janis Osis, Department of Applied Computer Science, Riga Technical University, Meza iela 1/3, Riga, LV 1048, Latvia; e-mail: Janis.Osis@cs.rtu.lv

of analyzing the system software developers set the main focus on analysis and modeling of software thus leading to a gap between the system and its supporting software [6]. This issue can be overcome by formalizing the very beginning of the software development lifecycle. By adding more efforts at the very beginning of lifecycle it is possible to build better quality software systems [7, 8].

By having too fuzzy beginning of the software development and lacking a good structure of it, the elimination of gap between computation independent viewpoint and the platform independent viewpoint depends much on designers' personal experience and knowledge (both viewpoints mentioned in the context of Model Driven Architecture – MDA [9]). Thus the quality of software system design models cannot be well controlled [10, 11]. There are a number of researches which try to enforce the initial phase in software development by strengthening it with various models like use cases [12], goal based models [13], behavioral models (like Activity and Sequence diagrams) [14], and structural models [15]. Previous researches in the field of formalizing very beginning of software development lifecycle propose TopUML modeling that enables modeling the functioning of both the problem and solution domains [16, 17]. Additionally it supports early solution domain model validation against functioning of the problem domain. TopUML modeling is a model-driven approach which combines Topological Functioning Model (TFM) [18] and its formalism with elements and diagrams of TopUML [19] (a profile based on UML). The TFM holistically represents a complete functionality of the system from the computation independent viewpoint [20]. It considers problem domain information separate from the solution domain information.

The purpose of this research is to strengthen the TopUML modeling with formal development of State diagram thus enabling transformation from TFM to it and eliminating the gap between problem domain model and software design (solution) model. Thus the paper is organized into following sections. Section 1 discusses the UML modeling driven methods that supports analysis of object state transitions and composition of corresponding State diagrams. Section 2 explores TopUML modeling and the prerequisites that should be satisfied in order to formally develop State diagrams in strong relevance with the problem domain. This section gives the formal method of developing State diagram based on TFM, i.e., the TFM to State diagram transformation pattern. Section 3 shows an example of using functional characteristics to analyze state changes of objects based on enterprise data synchronization system. Paper is concluded with conclusions of the performed research.

1. Related Works

UML is a notation and as such its specification does not contain any guidelines of software development process (e.g., which diagrams to use in which order). In fact this is pointed out as one of the UML weaknesses [21]. Despite that UML is independent of particular methods and approaches, most of the UML modeling driven methods use Use Case driven approach [22]. This might be caused by the originators (Booch, Rumbaugh, and Jacobson) of the UML since they recommend a Use Case driven process in their book "The Unified Modeling Language User Guide" [23].

According to [24] a successful software development project can be measured against the deliverables that satisfy and possibly exceed expectations of customer, the delivery schedule that has occurred in a timely and economical fashion, and the created result is resilient to change and adaptation. For software development project to be successful by means of given measurements, it should satisfy the following two characteristics:

- Solution should have a strong architectural vision, and
- A well-managed development lifecycle should be used.

This section discusses the current state of the art of UML based software development approaches by paying attention on one aspect – support of analysis for object state changes and transitions:

- State diagrams within Unified software development process [2] are developed during elaboration and construction phases. The use of state diagrams is emphasized for showing system events in use cases, but they may additionally be applied to any class.
- Business Object-Oriented Modeling (B.O.O.M.) developed by Podeswa [1] states that at least for every key business object a state diagram should be created.
- According to GRASP patterns (General Responsibility Assignment Software Pattern) introduced by Larman in [25] the State diagrams are used to describe allowed sequence of external system events that are recognized and handled by a system in the context of a use case. Additionally state diagrams can be applied to any class.
- Conceptual modeling described in [26] states that each entity type may be associated with zero, one, or more State diagrams. Conceptual modeling can be viewed as an activity related to capturing the knowledge about the desired system functionality.
- State diagrams within Component based development are used to determine the threads of control within the system [27].

The reviewed methods share common viewpoint of the application of State diagrams within software development process:

- State diagrams are developed by analyzing Use cases (more precisely: the scenario described by it),
- One state diagram per class or object, and
- State diagram should be developed for each most important object within the system.

Above mentioned three statements regarding application of State diagrams raise a set of ambiguousness and questions. The Use cases cannot be considered as a complete problem domain representation and a formal connection between problem domain and the solution [28]. The application of Use cases to develop diagrams of other types (such as State diagram) depends much on the designers' personal experience and knowledge, thus leaving the following question open:

- How to formally eliminate and overcome the gap between problem domain model and the design models?, and
- What are "most important objects" and how to formally identify them?

To overcome these issues the TopUML modeling is applied within software development as described in the next section.

2. Object State Change and Transition Analysis by using Functional Characteristics of Problem Domain

The object state change and transition analysis by using functional characteristics of problem domain is based on TopUML modeling, which is a model-driven approach intended to model problem domain and design software systems [19]. It combines TFM and its formalism with elements and diagrams of TopUML. The TFM considers problem domain information separate from the solution domain information and holistically represents a complete functionality of the system from the computation independent viewpoint while TopUML has elements of representing system design at the platform independent viewpoint and platform specific viewpoint. TFM has strong mathematical basis and is represented in a form of a topological space (X, Θ), where X is a finite set of functional features of the system under consideration, and Θ is the topology that satisfies axioms of topological structures and is represented in a form of a directed graph [18].

The application of TopUML modeling ensures proper analysis of system functioning by identifying and analyzing the functioning cycles. By using TopUML the information of system functioning from TFM can be transferred to design models thus allowing marking and evaluating the most important objects and components within system and to assign proper responsibilities to the right objects in a formal way. The most important objects are the ones that are participating in the main functioning cycle of the system. The main functional cycle is a directed closed loop that shows the functionality of system which is essential to its existence. By interrupting the main functional cycle the system cannot function or even exist. [19, 29]

In the context of state change analysis of objects the following TopUML modeling activities should be performed:

- TFM development (see Section 2.1 below),
- Domain model analysis and design (see Section 2.2), and
- Object state change and transition analysis (see Section 2.3).

2.1. Topological Functioning Model Development

The development of TFM consists of four steps. By completing these steps a TFM representing complete functioning of the problem domain gets developed. Afterwards the TFM is used as a source for development of other diagrams thus overcoming the gap between problem and solution domains. The four steps of TFM development are as follows:

Step 1: Definition of physical or business functional characteristics which consists of the following activities [30]:

1. Definition of objects and their properties from the problem domain description;
2. Identification of external systems and partially-dependent systems; and
3. Definition of functional features using verb analysis in the problem domain description, i.e., by finding meaningful verbs.

As a result a set of functional features are defined. Each functional feature X_{id} is a unique tuple specifying an action in problem and solution domains [18]. Equation (1) defines tuple of functional feature together with elements required to transform TFM into State diagram:

$$X_{id} = <Id, A, Op, R, O, Cl, St, PrCond, PostCond, E, Es, S>, \text{ where} \tag{1}$$

- *Id* – identifier of functional feature,
- *A* – action of object *O*,
- *Op* – operation which will provide functionality defined by action *A* (can be acquired when the class diagram is synthesized),
- *R* – result of action *A* (optional),
- *O* – object that receives the result or that is used in action *A* (for example, a role, a time period, a catalogue, etc.),
- *Cl* – class which will represent object *O* in static viewpoint of system (can be acquired when the class diagram is synthesized),
- *St* – new state of object *O* after performing action *A* (optional),
- *PrCond* – is a set of preconditions (optional),
- *PostCond* – is a set of postconditions (optional),
- *E* – entity responsible for performing action *A*,
- *Es* – indicates if execution of action *A* could be automated (i.e., performed without human interaction), and
- *S* – subordination of functional feature (can be internal or external).

At the lowest abstraction level one functional feature describes only one atomic action. Atomic action means that it cannot be further divided into a set of business actions. The functional features are represented as vertices in a directed graph of TFM.

Step 2: Introduction of topology Θ (in other words – creation of topological space) which involves establishing cause-and-effect relations between identified functional features. Cause-and-effect relations are represented as arcs of a directed graph that are oriented from a cause vertex to an effect vertex. Topological space represents the system under consideration together with the environment in which this system exists.

Step 3: Separation of TFM from topological space which is done by applying the closure operation over a set of system's inner functional features [31]. Initial TFM can be called "*TFM as-is*" where "*as-is*" means that the TFM represents the functioning of the problem domain without the impact of planned software system. Construction of initial TFM can be iterative. Iterations are needed if the information collected for TFM development is incomplete or inconsistent or there have been introduced changes in system functioning or in software requirements. The TFM development steps 1 to 3 can be partly automated as shown in [32] where the business use cases are automatically transformed into TFM.

Step 4: Identification of logical relations between cause-and-effect relationships consists of two steps since there are two kinds of logical relationships – one kind is between arcs that are outgoing from functional features and the other kind is between arcs that are incoming to functional features. Thus the identification of logical relations consists of two actions:

1. Identification of logical relations between cause-and-effect relationships that are outgoing from functional feature, and
2. Identification of logical relations between cause-and-effect relationships that are incoming to functional feature.

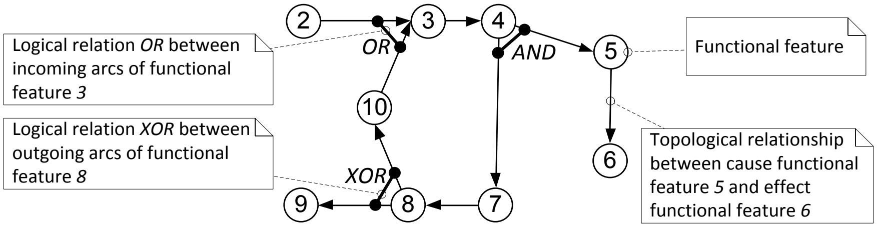

Logical relation *OR* between incoming arcs of functional feature *3*
Logical relation *XOR* between outgoing arcs of functional feature *8*
OR
AND
XOR
2
3
4
5
6
7
8
9
10
Functional feature
Topological relationship between cause functional feature *5* and effect functional feature *6*

Figure 1. Example of TFM

Each logical relation consists of two or more cause-and-effect relationships and a relation type. Within TFM can be defined three types of logical relations:

- Conjunction (*and*),
- Disjunction (*or*), and
- Exclusive disjunction (*xor*).

An example of TFM consisting of nine functional features, nine topological (i.e., cause-and-effect) relationships and three logical relations is given below in Figure 1.

2.1.1. Mappings between TFM and State Diagram

This section discusses the mappings between elements of TFM and State diagram. The mappings between standard UML diagrams can be found in various books and researches, like [1, 23, 27, 33] and [34]. Mappings between elements of TFM and State diagram are described in the form of table (see Table 1) by giving element of TFM and corresponding element in State diagram. For better understanding in addition a description of each mapping is given.

Table 1. Mappings between elements of TFM and elements of State diagram

TFM element	State diagram element	Description
Object state (*St* (1) specified by functional feature)	State	Each functional feature specifies an object performing certain operation. If during execution of this action changes the state of object performing this action, functional feature specifies the new state of the object. Object state from functional feature is transformed into state in State diagram.
	Initial state	When information from input functional feature is transformed into a state, an initial state is added before this state.
	Final state	When information from output functional feature is transformed into a state, a final state is added after this state.
Topological relationship	Transition	If during execution of action specified by functional feature is changed the state of object performing this action then incoming topological relationship defines transition from previous state to the new state.
Operation (*Op* (1) specified by functional feature)	Event	Each functional feature specifies an atomic business action which later is specified by topological operation in TFM. If functional feature specifies the new state of object, the operation is transformed into the event triggering transition from one state to another.
	Entry effect	If current functional feature specifies the new state of object, the operation is transformed into the entry effect of this new state.

TFM element	State diagram element	Description
	Exit effect	If descendant functional feature specifies the new state of object, the operation of this descendant functional feature is transformed into the exit effect of current state.
Preconditions of functional features (*PrCond* and *PostCond* (1))	Guard condition	If current functional feature specifies the new state of object, the preconditions of this functional feature are transformed into the guard conditions.
Logical relationship with type "*and*" (and partially "*or*")	Fork and Join	A logical relation in TFM give additional information about execution concurrency of functional features, thus conjunction (and) within State diagram is represented with fork and corresponding join. Disjunction (or) indicates of possible fork and join.

2.2. Domain Model Analysis and Design

Domain model analysis and design within TopUML modeling is based on the Topological class diagram and consists of the following two steps:

Step 1: Analysis of objects and their communication is based on the TFM transformation into Communication diagram (in previous researches the Problem domain objects graph was use instead of Communication diagram [19]). This transformation can be done automatically since TFM has all the information that is necessary for Communication diagram. When transforming TFM into Communication diagram the following are used:

- Functional features – source for lifeline identification and message sending from object to object,
- Topological relationships – determines the message sender and receiver as well as the message sending sequence, and
- Logical relations – shows the message sending concurrency.

In order to obtain a Communication diagram, it is necessary to check if each functional feature of the TFM reflects only one type of object. If some of functional feature reflects more than one type of object then it is needed to decompose it to the level where one functional feature uses only one type of objects. If TFM has been successfully checked it can be transformed into Communication diagram. The first step in transformation is to merge functional features with objects of the same type in one lifeline (the lifeline represents the class attribute of the functional feature). While merging functional features into lifelines the relationships with other lifelines should be retained (if there is more than one topological relationship then only one link is added between lifelines). Actors to Communication diagram are added from the input functional features.

For a better understanding of TFM to Communication diagram transformation, a small fragment of TFM consisting of two functional features A and B is used (see Figure 2), where A is an input functional feature of TFM and dashed arrows show mappings between elements of TFM and Communication diagram.

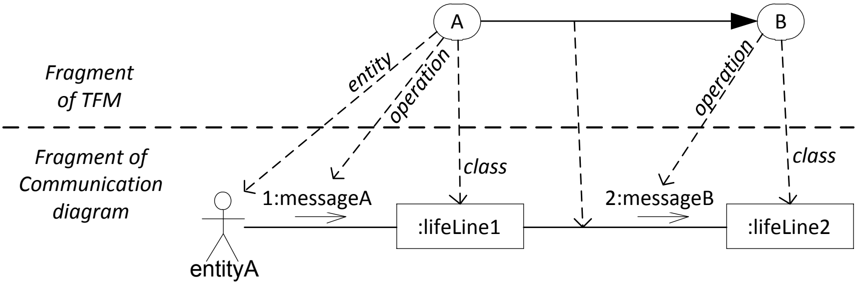


Figure 2. Example of TFM to Communication diagram transformation

Step 2: Domain model development by means of Topological class diagram consists of four activities:

1. Adding classes and operations,
2. Adding topological relationships between classes,
3. Identifying attributes, and
4. Refining initial Topological class diagram.

At first the Communication diagram is used for adding classes and operations to the Topological class diagram – lifelines are transformed into classes and messages into operations. The next step is adding topological relationships between classes. Since the notation of Topological class diagram allows variations of topological relationship graphical representation, it is advised to draw only one directed arrow in the same direction between classes (the arrow will show the cause and the effect operations).

After the classes and topological relationships between them have been established the next step is identification of attributes. This can be achieved by taking into consideration attributes of the object represented by functional feature. If the functional feature is well specified the class attribute of it is determined. If the class attribute is not determined, it can be specified in several ways (e.g. by analyzing functioning description of the system and searching nouns that represents attributes of the object [35], performing expert interviews [1], or by using ontology [36]).

By transforming Communication diagram an initial Topological class diagram is obtained (with attributes, operations, and topological relations between classes). A topological relation shows the control flow within the system. If static relations should be included (such as associations, generalization, etc.) then initial topological class diagram should be refined [37].

2.3. Object State Change and Transition Analysis

Object state change and transition analysis is based on the TFM transformation into a set of State diagrams (see Figure 3). The input of this activity is refined TFM and classes (either from Topological class diagram or lifelines from Communication diagram) and the output of this activity is one State diagram for each class.

Each functional feature specifies an object performing certain action. The count of obtained State diagrams is denoted by count of distinct objects specified by functional features. It is advised to analyze state changes of complex or most important objects in the system [1]. The most important objects are denoted by TFM – the functional features that are included into main functional cycle denote them, thus the identification of most important objects are done in a formal way.

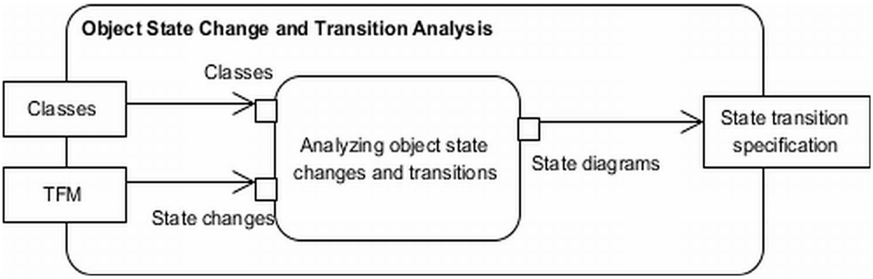

Object State Change and Transition Analysis
Classes
Classes
TFM
State changes
Analyzing object state changes and transitions
State diagrams
State transition specification

Figure 3. Analyzing object state changes and transitions

The first action is to scale down TFM which is performed by removing functional features which does not represent the object under consideration but in the same time retaining cause-and-effect relations. For example, assume that TFM consists of three functional features A, B, and C and are in the following causal chain: A→B→C. The A and C represent the same object while B represents another object. The resulting (scaled down) TFM is as follows: A→C.

States for each class are obtained from the functional features of refined TFM (functional feature has an attribute that defines the new state of the object). If the execution of functional feature involves the change of the corresponding object's state, then the state attribute has value, otherwise the value is not set. State transitions are obtained by transforming cause-and-effect relationship between functional features.

The special states (initial state and final state) are added to the obtained State diagram as follows:

- The initial state is added before the states that are obtained from the functional features which are the inputs of the downscaled TFM (the ones which has no predecessors), and
- The final state is added after the states that are obtained from the functional features which are the outputs of the downscaled TFM (the ones which has no descendants).

The example of transforming generic example of TFM into State diagram is given in Figure 4.

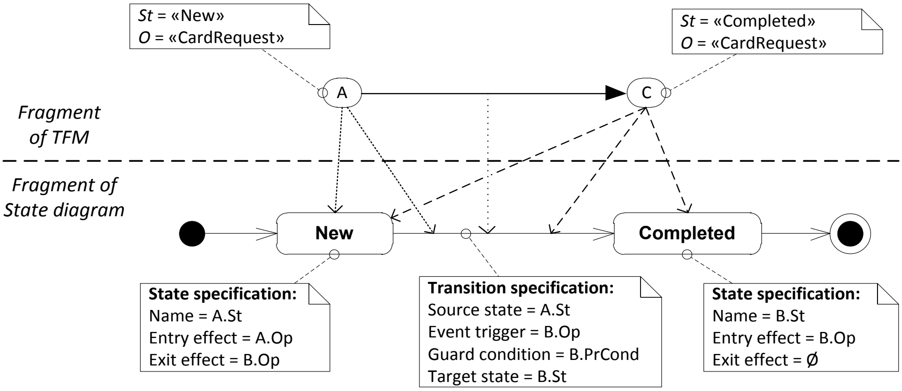


Figure 4. Example of TFM to State diagram transformation

3. Example of Object State Change and Transition Analysis

Example of object state change and transition analysis by using functional characteristics of problem domain is based on a case study in which TFM is developed for enterprise data synchronization system. The enterprise data synchronization system is developed by applying TopUML modeling and involves creation of TFM, Use case diagram, Problem domain objects graph (applied instead of Communication diagram), Topological class diagrams, and Sequence diagrams [16].

Within the case study have been defined 30 functional features by analyzing functioning of enterprise data synchronization system. Part of defined functional features is given in Table 2 where are included features that specify the new state for object named *"Scheduler"*. After definition of functional features the topology Θ (cause-and-effect relationships) is identified between those functional features thus creating topological space. In order to get the TFM the closuring operation is applied over the set of internal system functional features. The developed TFM after applying closuring operation is as follows: X={2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 24, 25, 26, 27, 28, 29}. The resulting graph is given in Figure 5 (a) which shows functional features (vertices), cause-and-effect relationships (arcs between vertices).

Table 2. Part of functional features defined for enterprise data synchronization system

ID	Object action (*A*)	Precondition (*PrCond*)	Object (*O*)	New state (*St*)
5	Reading all data from source data base	If import should be performed from source data base	Scheduler	Reading data
6	Checking if read data structure is according to specification		Scheduler	Checking data
7	Putting the read data into temporal internal table	If data structure is according to specification	Scheduler	Importing
9	Checking import folder		Scheduler	Reading data
12	Checking if import file data structure is according to specification		Scheduler	Checking data
13	Converting the read data from import file into temporal internal table	If import file structure is according to specification	Scheduler	Importing
15	Moving import file to processed files folder		Scheduler	Completing import
19	Checking if data from a particular row already exists in target data base		Scheduler	Importing
25	Logging data row from temporal internal table		Scheduler	Logging status
29	Archiving log file	If data import is completed	Scheduler	Completing import

The example of object state change analysis in the context of enterprise data synchronization system development case study is performed for the object name *"Scheduler"*. The functional features specification in Table 2 shows that this object in total has five different states: 1) *Reading data*, 2) *Checking data*, 3) *Importing*, 4) *Logging status*, and 5) *Completing import*. The resulting State diagram is given in Figure 5 (b).

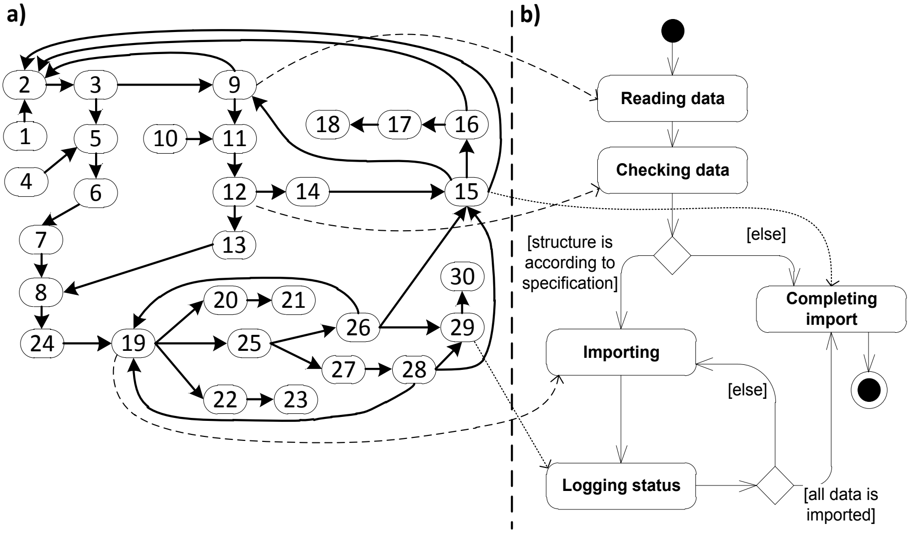


Figure 5. TFM of enterprise data synchronization system functioning (a) and State diagram for object *"Scheduler"* (b)

4. Conclusions

The main goal of this research is to do formal development of State diagram by analyzing functional characteristics of a problem domain. The result of research is method for transforming TFM into State diagram thus eliminating the gap between problem domain model and software design (solution) model.

UML modeling driven methods (like Unified process, B.O.O.M. and patterns based software development) manifests that the State diagrams are developed by analyzing Use cases (more precisely: the scenario described by it), one state diagram per class or object. In fact they say that State diagram should be developed for each most important object within the system. These statements raise a set of ambiguousness and questions. The Use cases cannot be considered as a complete problem domain representation and a formal connection between problem domain and the proposed solution. The application of Use cases to develop diagrams of other types (such as State diagram) depends much on the designers' personal experience and knowledge.

The elaborated TopUML modeling (including the State diagram development) proposes a way on how to formally overcome the gap between problem domain and solution domain – the first one is represented by TFM which shows the complete functioning of a problem domain and the latter one is obtained by transforming TFM of a problem domain. Moreover the TopUML enables formal identification of the most important objects and classes within system – they are denoted by TFM: functional features that are included into main functional cycle specify these objects and classes. In contrast, the reviewed UML modeling driven methods relies that the designers' personal experience and knowledge is sufficient to identify most important objects within system. In addition the example described in paper shows State diagram development for the case study in which enterprise data synchronization system has been developed by using TopUML modeling.

This research shows that by adding additional efforts at the very beginning of software development life cycle it is possible to create a model that contains sufficient and accurate information of problem domain. By “sufficient” meaning that this model can be transformed into other diagrams without major re-analysis of problem domain and by “accurate” meaning that the model precisely reflects the functioning and structure of the system.

Acknowledgement

This work has been supported by the European Social Fund within the project “Support for the implementation of doctoral studies at Riga Technical University”.

References

[1] H. Podeswa, *UML for the IT Business Analyst.* 2nd ed. Course Technology PTR, 2009.

[2] K. Scott, *The Unified Process Explained*. Addison-Wesley, USA, 2001.

[3] *Unified Modeling Language Superstructure version 2.3*. OMG, May 2010. Available from: http://www.omg.org/spec/UML/2.3/Superstructure/PDF/.

[4] C. Jones, Positive and negative innovations in Software Engineering, *International Journal of Software Science and Computational Intelligence* **1**(2) (2009), 20-30.

[5] J.Osis and E. Asnina, A business model to make software development less intuitive. In: *Proceedings of the 2008 International Conference on Innovation in Software Engineering*, Vienna, Austria. IEEE Computer Society CPS, Los Alamitos, USA, 2008, 1240-1246.

[6] J. Osis, Formal computation independent model within the MDA life cycle, *International Transactions on Systems Science and Applications* **1**(2) (2006), 159- 166.

[7] J. Osis and E. Asnina, Topological modeling for model-driven domain analysis and software development: functions and architectures. In: *Model-Driven Domain Analysis and Software Development: Architectures and Functions*, IGI Global, Hershey - New York, 2011, 15-39.

[8] J. Osis and E.Asnina, Is modeling a treatment for the weakness of Software Engineering? In: *Model-Driven Domain Analysis and Software Development: Architectures and Functions*, IGI Global, Hershey - New York, 2011, 1-14

[9] J. Miller and J. Mukerji, editors, *MDA Guide Version 1.0.1*. OMG, 2003.

[10] J. Osis, E. Asnina, and A. Grave, MDA oriented computation independent modeling of the problem domain. In: *Proceedings of the 2nd International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2007)*, Barcelona, Spain, 2007, 66 -71.

[11] W. Zhang, H. Mei, H. Zhao, and J. Yang, Transformation from CIM to PIM: A feature-oriented component-based approach. In: *Model Driven Engineering Languages and Systems*, Lecture Notes in Computer Science **3713** (2005), Springer-Verlag, Berlin, 248-263.

[12] T. Yue, L. Briand, and Y. Labiche, A use case modeling approach to facilitate the transition towards analysis models: concepts and empirical evaluation. In: *Model Driven Engineering Languages & Systems*, Notes in Computer Science **5795** (2009), Springer-Verlag, Heidelberg, 484-498.

[13] E. Letier and A. van Lamsweerde, Deriving Operational Software Specifications from System Goals. In: *Proceedings of the 10th ACM SIGSOFT Symposium on Foundations of Software Engineering*, ACM, New York, 2002, 119-128.

[14] I. Diaz, O. Pastor, and A. Matteo, Modeling interactions using role-driven patterns. In: *Proceedings of 13th IEEE International Conference on Requirements Engineering (RE 2005)*, Paris, France. IEEE Computer Society, 2005, 209-220.

[15] E. Insfran, O. Pastor, and R. Wieringa, Requirements engineering-based conceptual modeling, *Requirements Engineering* **7**(2) (2002), 61-72.

[16] U. Donins, and J. Osis, Topological modeling for enterprise data synchronization system: a case study of topological model-driven software development. In: *Proceedings of the 13th International Conference on Enterprise Information Systems (ICEIS 2011)* **3** (2011), SciTePress, 87-96.

[17] U. Donins, Software development with the emphasis on topology. In: *Advances in Databases and Information Systems*, Notes in Computer Science **5968** (2010), Springer-Verlag, Berlin, 220-228.

[18] J. Osis and E. Asnina, *Model-Driven Domain Analysis and Software Development: Architectures and Functions*, IGI Global, Hershey-New York, USA, 2011.

[19] J. Osis and U. Donins, Platform independent model development by means of Topological Class Diagrams. In: *Model-Driven Architecture and Modeling Theory-Driven Development*, 2nd international MDA & MTDD workshop in conjunction with ENASE 2010, Athens, Greece, SciTePress, Portugal, 2010, 13-22.

[20] J. Osis and E. Asnina, Enterprise modeling for information system development within MDA. In: *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008),* USA, 2008, 490.

[21] S. Kent, The Unified Modeling Language. In: *Formal Methods for Distributed Processing: A Survey of Object-Oriented Approaches*, 1st edition (October 22, 2001), Cambridge University Press, 126-151.

[22] B. Dobing and J. Parsons, Dimensions of UML diagram use: practitioner survey and research agenda. In: *Principle Advancements in Database Management Technologies: New Applications and Frameworks*, IGI Global, 2010, 271-290.

[23] G. Booch, J. Rumbaugh, and I. Jacobson, *The Unified Modeling Language User Guide*. 2nd ed. Addison-Wesley, 2005.

[24] G. Booch, R. Maksimchuk, M. Engel, B. Young, J. Conallen, and K. Houston, *Object-Oriented Analysis and Design with Applications*. 3rd ed. Addison-Wesley, 2007.

[25] C. Larman, Applying UML and Patterns: *An Introduction to Object-Oriented Analysis and Design and Iterative Development*. 3rd ed. Prentice Hall, 2005.

[26] A. Olive, *Conceptual Modeling of Information Systems*, Springer, 2007.

[27] P. Stevens and R. Pooley, *Using UML: Software Engineering with Objects and Components*. 2nd ed. Addison-Wesley, (2005).

[28] J. Osis and E. Asnina, Derivation of use cases from the topological computation independent business model. In: *Model-Driven Domain Analysis and Software Development: Architectures and Functions*, IGI Global, Hershey - New York, 2011, 65-89.

[29] J. Osis, E. Asnina, and A. Grave, Formal problem domain modeling within MDA. In: *Evaluation of Novel Approaches to Software Engineering*. Communications in Computer and Information Science **22** (2008), Springer-Verlag, Berlin, 387-398.

[30] J. Osis, E. Asnina, and A. Grave, Formal computation independent model of the problem domain within the MDA. In: *Proceedings of the 10th International Conference on Information Systems and Formal Models (ISIM'07)*, Silesian University in Opava, Czech Republic, 2007, 47-54.

[31] E. Asnina and J. Osis, Topological Functioning Model as a CIM-business model. In: *Model-Driven Domain Analysis and Software Development: Architectures and Functions*, IGI Global, Hershey - New York, 2011, 40-64.

[32] J. Osis and A. Šlihte Transforming textual use cases to a computation independent model. In: *Proceedings of the 2nd International Workshop on Model-Driven Architecture and Modeling Theory-Driven Development*, Athens, Greece, July 22-24, 2010, 33-42.

[33] O. Nikiforova, Object interaction as a central component of object-oriented system analysis. In: *Proceedings of the 2nd International Workshop on Model-Driven Architecture and Modeling Theory-Driven Development*, Greece, Athens, July 22-24, 2010, 3-12.

[34] R. Van Der Straeten, T. Mens, J. Simmonds, and V. Jonckers, Using description logic to maintain consistency between UML models. In: *UML 2003 - The Unified Modeling Language, Modeling Languages and Applications*. Lecture Notes in Computer Science **2863** (2003), Springer-Verlag, Berlin, 326-340.

[35] J. Osis and U. Donins, Formalization of the UML class diagrams. In: *Evaluation of Novel Approaches to Software Engineering*. Communications in Computer and Information Science **69** (2010), Springer-Verlag, Berlin, 180-192.

[36] X. Li and J. Parsons, Ontological semantics for the use of UML in conceptual modeling. In: *Challenges in Conceptual Modelling. Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling*, Auckland, New Zealand, November 5-9, 2007, 179-184.

[37] U. Donins, J. Osis, A. Šlihte, E. Asnina, and B. Gulbis, Towards the refinement of topological class diagram as a platform independent model. In: *Proceedings of the 3rd International Workshop on Model-Driven Architecture and Modeling-Driven Software Development*, China, Beijing, June 8-11, 2011, 79-88.

Electronic Archive Information System

Saulius RAGAISIS [a,1], Adomas BIRSTUNAS [b], Antanas MITASIUNAS [b] and Arunas STOCKUS [b]
[a] *Software Engineering Department, Vilnius University, Lithuania*
[b] *Computer ScienceDepartment, Vilnius University, Lithuania*

Abstract. Development of Electronic Archive Information System was the final step in completing Lithuania's preparation for signed electronic documents. The system is unique not only in Lithuania, but also in Europe. It allows to submit to the state archives the official electronic documents signed by qualified e-signature, assuring their integrity, authenticity, non-repudiation and possibility to use and store them for a long or unlimited time. To resolve the interoperability problems the universal model, sufficient for complex electronic documents during whole of their life cycle has been created and described in two specifications: ADOC - for human readable electronic documents and MDOC - for machine readable electronic documents. The public free software tools for preparation, signing, preview and verification of official electronic documents have played an important role in promoting the usage of electronic documents. The paper presents practical fundamental solutions applied in Lithuania.

Keywords. State archives, official electronic documents, qualified e-signature

Introduction

Life in the digital age raises the need for electronic documents. E-Government, e-business, e-commerce, and other e-services are impossible without electronic documents and security features that must be based on PKI (Public Key Infrastructure) technology [1].

Paper documents are no longer sufficient. If the original document is paper based, usually it is scanned to save time and costs of ordinary post services. However, detailed calculation shows that in the case of Lithuania the scanning costs per year exceed 10 M€. In European Union such procedure costs few billion yearly. Use of official electronic documents also provides a lot of other advantages: full reliability of document content, possibility to use fragments of documents for preparation of other documents, possibility to search in a document text. Quality of electronic documents is certainly better than that of scanned paper copies. So, it is completely clear that with time, electronic documents will replace conventional documents everywhere.

First Lithuania wide system with e-signature was the e-Servicing System of the Insurers (EDAS) launched by State Social Insurance Fund Board of the Republic of Lithuania (SoDra) in the end of 2007. It allows the insurers (employers) to provide digitally signed Abbyy eFormFiller forms instead of hand-signed paper documents. As

[1] Corresponding Author: Software Engineering Department, Faculty of Mathematics and Informatics, Vilnius University, Naugarduko Str. 24, LT-03225 Vilnius, Lithuania; E-mail: saulius.ragaisis@mif.vu.lt.

these forms are XML based, XAdES enveloped signature format has been selected as the simplest but suitable for this application approach.

The main issue with the electronic document usage is the interoperability. At the moment, most of the EU member states are still in the process of defining clear strategy in the area of the official electronic documents and their usage in every day communication between businesses, citizens and public sector. The main standardization efforts in EU are still targeted to electronic signature formats. Decision 2011/130/EU [2] obliges member states to be able to technically process some of the most common formats of advanced e-signatures XAdES, CAdES, and PAdES. But the true artefact of interoperability is an electronic document. Lithuanian solution for the interoperability problems was the universal model, sufficient for complex electronic documents during whole of their life cycle that was defined in specifications ADOC and MDOC which were approved by Office of the Chief Archivist of Lithuania (Lithuanian Archives Department until year 2011).

The public free software tools for preparation, signing, preview and verification of official electronic documents according to the approved specifications have played an important role in promoting the usage of electronic documents. These tools are available as Web-based and desktop applications.

Currently more than half million machine-readable electronic declarations signed by qualified electronic signature are created each month using systems of State Social Insurance Fund Board and State Tax Inspectorate. The ministries and state main institutions have upgraded their document management systems with functionality for creating and verifying official electronic documents.

Another important date on the road to electronic documents is September 7, 2011, when Lithuania's Prime Minister and the cabinet have put their first electronic signatures to the legal acts passed by the Government. Since then, the originals of Government decrees are electronic documents of ADOC format. The electronic signature system ELPAS allows the Government to manage documents more efficiently by allowing to electronically sign the submitted legal acts at any convenient time and in any place. The Government also uses ELPAS for the submission of draft Presidential decrees as well as draft laws to the Seimas.

Over recent years there has been an increase in awareness on the part of state archives of the EU Member States of the need to preserve electronic records. The legal situation across Europe with regard to the archiving of public records generated by government agencies, and public access to such records, remains diverse. Differences exist not only in terms of the legal requirements, but also in the extent to which these are interpreted and enacted [3]. Development of Electronic Archive Information System (EAIS) was the final step in completing Lithuania's preparation for signed electronic documents. EAIS is intended for efficiently administering the whole National Document Fund (NDF), including paper documents, and providing electronic services. The system is unique not only in Lithuania, but also in Europe because the electronic documents it deals with are original documents, having the same legal value as handwritten documents. EAIS assures the integrity, authenticity, non-repudiation and possibility to use the electronic documents for a long or unlimited time.

1. Model of Electronic Document

The electronic documents' interoperability issue influences even the understanding of an electronic document as such. Interoperability is ensured if the electronic document is created and accepted using the same document processor. Therefore, the document processor's format often is treated as format of electronic document. Lithuanian approach is based on the principle that an electronic document as independent entity should be adequate to the conventional documents.

A real document is very complex entity. The document provides some content that may consist of texts, spreadsheets, tables, calculations, drawings, pictures, etc. as well as appendices or attached independent earlier created documents. In turn the appendices may contain their own appendices. An attached document may contain other attached documents and so on. The documents having legal value must be signed by one or more signatures. A signature expresses the conscious and voluntary will of the signatory to approve the content signed "lu et approuvé". Official documents achieve legal value only after signed documents are registered and possess corresponding attributes. Therefore, in general the documents consist of three parts: content, signatures and metadata.

An adequate electronic document having legal value equal to handwritten documents should be very complex entity too. It also must consist of content, signatures and metadata. However, usually these three areas are addressed by three different fields of interests – document processors, electronic signatures initiatives and record management systems, i.e. file formats, electronic signatures formats and metadata are standardized separately but signed electronic document format is not standardized.

An electronic document has a complex life cycle starting with document creation, including sophisticated document creation procedure, usage, storage, until destruction if a document has limited life cycle.

First of all, in October 2008 Lithuanian Archives Department has adopted the minimal requirements for the specifications of electronic document signed by the electronic signature that addresses all three components of the document: content, signatures and metadata without limitation of complexity of document's structure. The task was to create universal electronic document model to be sufficiently powerful to address the needs for complex electronic documents operated by public sector institutions including documents' whole life cycle. In this context, the natural choice without alternatives was a zip-based electronic documents format approach. The document format (container) is conformant with the requirements for signature containers defined by ETSI TS 102 918 [4].

The requirements for content are defined separately at logical and physical levels. The logical level operates by such notions as main document, one or more appendices, and one or more attached independent documents. The physical level is defined in terms of files and directories including file formats allowed.

XAdES electronic signature format of detached topology shall be applied according to ETSI TS 101 903 [5]. Multiple signed data objects and multiple signatures are allowed. Metadata as a sub-tree of xml-based metadata file can be signed.

The basic principle of Lithuanian approach is that metadata is an integral part of the electronic document. International standard ISO 15489-1 [6] indicates that "the metadata embedded in, attached to, or associated with, a specific record". This means that according to standard, metadata can be embedded into a document or attached to a document, or associated with a document. Electronic documents interoperability issues and the needs for electronic documents automated processing force to use embedded metadata as an integral part of electronic document.

Currently there are two main official electronic documents' specifications adopted by Lithuanian Archives Department: ADOC [7] – for signed human readable electronic document and MDOC – for machine-readable electronic documents. These specifications were awarded as the main strategic innovation 2010 in Lithuania.

It could be noted that e-Servicing Systems of SoDra are based on different specifications:

- the Citizens system (launched in 2009) on EGAS specification that is subset of MDOC specification;
- the Insurers system (launched in 2007) on EDAS specification that is not conformant with minimal requirements for the specifications.

Universal electronic documents model implemented by specifications ADOC and MDOC has sufficient power to express such extremely complex documents as State Budget 2012.

2. Implementation of Electronic Documents Specification

The electronic documents specification defines the format of valid electronic document but not the procedure how such format should be produced. The software requirements for creation and verification of electronic documents are defined in the international technical standards [8, 9, 10, 11].

Most solutions in Lithuania with qualified e-signature according to ADOC specification are based on product line Signa. It consists of four products:

- Signa Desktop – public free Windows OS application (available to download from http://www.mitsoft.lt/);
- Signa Web – public free web application (available at https://signa.mitsoft.lt/);
- Signa SDK – a set of application libraries for Java and .NET platforms; it supports MDOC and EGAS specifications also;
- Signa Docs – web application for enterprises featuring multiple users, electronic document workflows and the ability to sign with qualified signatures multiple electronic documents in bulk (ELPAS is based on it).

Signa is known to provide the must complete implementation of ADOC specification, from documents creation to documents validation, e-signatures verification, signature formats upgrade. Both signing possibilities – using local secure signature creation device as well as mobile electronic signature services – are ensured. Product line Signa was awarded as a winner of national contest "Innovative Product 2011".

3. Archive of Electronic Documents

The goal of the Electronic Archive Information System (EAIS) project was to create an integral open information system for accepting and storing electronic documents of

National Document Fund (NDF), providing a legal access to the stored documents using IT and communication means, administering NDF efficiently and providing electronic services. Experience of other countries [12, 13] as well as research ideas [14, 15] has been taken into account.

EAIS consists of the 3 main parts:

- Public portal (https://eais-pub.archyvai.lt) that serves all external users;
- Internal portal (https://eais-int.archyvai.lt) that serves the users of state archives and Office of the Chief Archivist of Lithuania;
- Storage of electronic documents.

Physically e-documents are stored in two geographically remote electronic archive data centers: one in Vilnius, another in Šiauliai. The system implements the replication of archive data between the main and the reserve data centers with the possibility to switch operations between the centers in case of a failure. Because of security reasons, the storage of electronic documents is accessible through internal portal only.

Authentication of external users is implemented through e-government gateway. The authentication service is provided for the users of Internet banking systems of all commercial banks operating in Lithuania and owners of class 2 or 3 personal digital certificates. It should be noted that some EAIS functions (e.g. creation of electronic documents, search in NDF) are available for not authenticated users also.

EAIS consists of the following modules:

- Acceptance of electronic documents (both portals);
- Retention of electronic documents (internal portal only);
- Publication and presentation of documents (both portals);
- Organization of documents management (both portals);
- Administration (both portals);
- Software tools of free accessibility for preparation, signing, preview, and verification of official e-documents (public portal only);
- Accumulation and analysis of statistical data about the stored documents and their usage (both portals).

It is important to emphasize that all functions of the public portal could be invoked interactively and through corresponding Web services. It is intended that organizations will extend functionality of their document management systems and operate with EAIS through Web services.

Acceptance of electronic documents module provides functionality to agree the suitable time for electronic documents transfer to state archive, to transfer the documents package through computer network or load it from physical media, to perform checks of e-documents integrity, authenticity and conformance to specification, to prepare them for storage a long or unlimited time, and to store.

The very important feature of EAIS is flexible configuration of authenticity checks. It is essential because some institutions may require final one signature over whole document to ensure document's integrity. Other institutions may require the use of qualified electronic signatures for any purpose of electronic signatures. Third institutions may require electronic document content signed by the purpose of signature, confirmation or approbation. All these partial cases are examples of valid electronic documents according to some specification.

The possibility to use electronic documents for unlimited time is assured by converting their contents into long-term storage files (PDF/A format) and formats for previewing the documents in Internet (PNG and JPEG). In the future these formats will be regularly reviewed and updated. The e-signatures that legal value (integrity, authenticity, non-repudiation) should be preserved are extended to the XAdES-A format.

Retention of electronic documents module includes the means for physical preservation of electronic documents, e.g. backup copies, saving original packages of electronic documents into WORM (write once, read many) devices, and the risk management. There are two types of risk identified:

- Related to content formats: over the time they could become old and not supported by current version of software;
- Related to electronic signatures: the cryptographic algorithms that are used for creating secure signatures today might become breakable in the future due to sudden advances in cryptoanalysis or in computational capabilities; the keys used for signing could become too short; the validity period of certificate used last time stamp could end or even certificate could compromised.

Risk of the first type is resolved by creation/renewing the long-term storage and previewing copies: all content files are transformed into the PDF/A format which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files [16]. Additional time stamp to XAdES-A format resolves risk of the second type. Of course, suitable up-to-date algorithms and the length of key should be used for this time stamp.

Archives of Lithuania are open to public. Any person, upon the presentation of an ID card, may have an access to archival records. Records of state institutions without restrains (except a few, access to which is restricted by law) are available for reference. Publication and presentation of documents module ensures the access to documents of the National Document Fund stored in EAIS. A person who wishes to get acquainted himself with the documents the access to which is not limited restricted, must authenticate through e-government gateway.

For the transparency and accountability purposes of the public administration state archives service play an important role in the processes of records management in public sector. State archives controls approximately 2000 institutions and agencies. State and municipal institutions, agencies and enterprises which transfer documents to state archives shall co-ordinate with the state archives the documentation plans, registers, and other registration documents of activity documents. Organization of documents management module allows to perform these functions by IT and communication means.

EAIS includes the functionality for providing applications for attested copies of documents, extracts intended for the approval of juridical facts, and certificates, management of these applications, preparation the copies of electronic documents and signing them.

The software tools of free accessibility for preparation of official e-documents, signing them by e-signature, preview, and verification include:

- Web application (available at https://adoc.archyvai.lt); if compared to Signa Web this application is more beginners oriented, as all the time user is guided by wizards;
- Signa Desktop.

EAIS has been launched in October 2011. It is the first national archive storing electronic documents signed with qualified e-signatures. Currently EAIS deals with electronic documents of ADOC and EGAS specifications.

4. Future Plans

There are two main trends in electronic documents formats evolution that can be called zip-based and pdf-based electronic documents formats. All specifications currently adopted in Lithuania define zip-based electronic documents format. Traditionally a portable document format (PDF) [17] is applied for documents of the flat structure. However, the evolution of documents structure has forced to consider it as a container format. Furthermore, in order to address the needs for extended functionality of signatures ETSI introduced PAdES signature format as new semantics based on PDF syntax.

It is possible to implement universal document model using pdf-based container but zip-based approach is more suitable for this purpose. However, in the case of simple documents PDF has some advantages. The document could be previewed with freely available PDF readers.

Therefore, PDF-LT specification is on the final stage of preparation. Correspondingly the minimal requirements for the specifications will be adjusted. PDF-LT specification will be recommended for short term documents only. But such documents could become long term or even unlimited term. Therefore, EAIS should be extended to accept and store electronic documents of PDF-LT specification.

Current version of the Law on Documents and Archives [18] defines that only permanent retention documents are accepted to state archives. Amendments of the legal base that long storage electronic documents could be submitted to the state archives also are prepared. They should be adopted by the Seimas of the Republic of Lithuania.

5. Conclusions

Lithuania's experience has shown that the “tool-first” approach does not lead to interoperability of electronic documents. Instead an opposite approach “standard-first” should be chosen. Implementing it, the minimal requirements for electronic documents specification and the specifications themselves – ADOC for unstructured human readable and MDOC for structured machine readable signed electronic documents – were developed based on universal document model to support all life cycle of the document. These specifications were nominated as main strategic innovation 2010 in Lithuania.

According to specification various interoperable tools for electronic documents creation and verification were developed and rolled-out for use of inhabitants, civil servants, providing public services. Documents management systems were integrated with functionality of signed electronic documents. Thanks to well established and now generally accepted document format the true interoperability between document management systems is about to be achieved: documents created and signed in one system may be opened in another system. Product line Signa was awarded as a winner of national contest "Innovative Product 2011".

The Lithuanian Government started to prepare and issue Decrees originals in electronic form.

Electronic Archive Information System is ready to accept electronic documents into the Lithuanian National Documentary Fund, to preserve their integrity, authenticity, non-repudiation for unlimited time, and to make them accessible to the citizens and governmental and public institutions, providing public services. The system is unique not only in Lithuania, but also in Europe.

The authors have played key role in development of all specifications, information systems and tools mentioned in the paper.

References

[1] P. Risztics and I. Jankovits, Electronic government and public administration in Hungary. In: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences HICSS05 Track 5,* 2005, 122a (1-8).

[2] Commission Decision 2011/130/EU of 25 February 2011 establishing minimum requirements for the cross-border processing of documents signed electronically by competent authorities under Directive 2006/123/EC of the European Parliament and of the Council on services in the internal market, *Official Journal of the European Union* **L 53** (26.2.2011), 66-72.

[3] I. Petravičiūtė, Long term preservation policy of electronic records in the European Union Countries, *Knygotyra* **47** (2006), 243-261 (in Lithuanian).

[4] *ETSI TS 102 918 V1.1.1: 2011-04 Electronic Signatures and Infrastructures (ESI); Associated Signature Containers (ASiC)*, European Telecommunications Standards Institute, 2011.

[5] *ETSI TS 101 903 V1.4.1:2009 XML Advanced Electronic Signatures*, European Telecommunications Standards Institute, 2009.

[6] *ISO 15489-1:2001 Information and Documentation - Records Management - Part 1: General*, International Organization for Standardization, 2001.

[7] Specification ADOC-V1.0 of the electronic document signed by the electronic signature, *Valstybės žinios* **108-4574** (09.10.2009) (in Lithuanian). Available from: https://signa.mitsoft.lt/static/signa-web/webResources/ docs/ADOC_specification_approved20090907_EN.pdf (in English).

[8] *CWA 14170:2004 Security Requirements for Signature Creation Applications*, European Committee for Standardization, 2004.

[9] *CWA 14171:2004 General Guidelines for Electronic Signature Verification*, European Committee for Standardization, 2004.

[10] *ETSI TS 101 861 V1.2.1:2002 Time Stamping Profile*, European Telecommunications Standards Institute, 2002.

[11] *ETSI TS 101 862 V1.3.3:2007 Qualified Certificate Profile*, European Telecommunications Standards Institute, 2007.

[12] *EDRMS Functional Specification, Standard*, Government of South Australia, Adelaide, 2009.

[13] *Functional Specification for Integrated Document and Records Management Solutions*, National Archives and Records Service of South Africa, Pretoria, 2004.

[14] A. J. Blazic, Long term trusted archive services. In: *ICDS '07 Proceedings of the First International Conference on the Digital Society,* IEEE Computer Society, 2007, 29-35.

[15] T. Kunz, S. Okunick, and U. Viebeg, Long-term security for signed documents: services, protocols, and data structures. In: *Long-Term and Dynamical Aspects of Information Security: Emerging Trends in Information and Communication Security*, Nova publishers, 2007, 125-139.

[16] *PDF/A – the Standard for Long-Term Archiving*, PDF Tools AG, 2009.

[17] *ISO 32000-1:2008 Document Management. Portable Document Format. Part 1: PDF 1.7*, International Organization for Standardization, 2008.

[18] *Law on Documents and Archives 18 June 2010 XI-917*, Seimas of the Republic of Lithuania, 2010. Available from: http://www3.lrs.lt/pls/inter3/dokpaieska.showdoc_l?p_id=404607.

Performance Measurement Framework with Indicator Life-Cycle Support

Aivars NIEDRITIS, Janis ZUTERS and Laila NIEDRITE[1]
University of Latvia

Abstract. The performance measurement method introduced in this paper is based on the five-step indicator lifecycle that covers formal definition of indicators, measurement, analysis, reaction, and reformulation of indicator definitions. Performance measurement framework is introduced that support this performance measurement method and that enables the indicator lifecycle. The goal of this research is to provide a method for performance measurement that ensures timely and to given context appropriate decision making process. For the purposes of storing the information necessary for decision support a data warehouse is used as a component of the process measurement framework.

Keywords. Performance measurement, key performance indicators, indicator lifecycle, data warehouse

Introduction

Effective organization of business processes ensures the achievement of institution's goals. Performance measurement compares the measurement results with target values to discover the progress. An important aspect is how to choose appropriate measures and how to define an appropriate measurement framework.

Performance measures [1] are indicators used by management to measure, report and improve the performance in an organization. What kind of particular performance measures are used is influenced by management models of organizations and measurement perspectives of these models. For example, BSC [2] defines four measurement perspectives: Financial, Customer, Internal Process, and Learning and Growth, other approaches add more perspectives, for example, Environment/Community and Employee Satisfaction [1].

To perform effective measurement and adequate reaction on discovered situations, not only different perspectives, but also different aspects (e.g. connection to success factors, reporting, reaction, responsibilities) of performance indicators could be modeled and documented. In our previous research [3] we investigated the features of indicators and grouped them according to the indicator life-cycle. This concept helps to support appropriate usage of indicators according the values of the features.

To implement performance measurement according to the strategies of the company and some management model, the companies develop and use measurement systems. A data warehouse is an option to build a Performance Measurement System.

[1] Corresponding Author: Laila Niedrite, University of Latvia, Raina bulv. 19, LV-1586, Riga, Latvia E-mail: Laila.Niedrite@lu.lv

An advantage of using a data warehouse for the implementation of a performance measurement system is the possibility to use existing infrastructure of the company's data warehouse.

Traditionally data warehouses store customer and financial indicators of the companies, but other perspectives are typically not covered. Some attempts to integrate the perspective of internal business processes into a data warehouse have been made in [4, 5, 6].

We do not try to incorporate another perspective of measurements into the data warehouse. We propose to use the data warehouse as an integral part of a performance measurement system, which can store indicators of different perspectives and can be used according to proposed measurement framework.

The measurement framework describes different measurement aspects to bring order within this important undertaking of the organization. Thereby, the quality of measurement is improved, for example, by means of performing the analysis of right indicators in right time and undertaking the right actions as a result.

The usage of an existing data warehouse gives additional advantage to the performance measurement. The analysis of indicator values can be performed using existing OLAP tools, reports and dashboards.

We start with related work described in section 1. Section 2 explains the concept of indicators life-cycle that forms the basis for proposed measurement framework. In section 3 the reporting tool and its metadata is described that is one of the ready-made data warehousing components used within the measurement framework. The architecture of performance measurement system is given in section 4. In section 5 the conclusions are given.

1. Related Work

Performance measurement systems implemented by means of a data warehouse are given in several works. The existing approaches concentrate mostly on how to build an appropriate dimensional model of the data warehouse according the process perspective of measures to be stored.

The Process Data Warehouse [5] stores histories of engineering processes and products for experience reuse. The Performance Management System (PMS) [6] stores financial and non-financial data centrally. The PMS contains values of measurements as well as supplementary information about company structure, business processes, goals and performance measures. Besides traditional data warehousing perspectives the process perspective is also analyzed. In [4], the authors propose a Corporate Performance Measurement System (CPMS), where process performance data is integrated with institution's data warehouse. Log files of a workflow system are used as data sources. The model of CPMS is developed as a part of an existing data warehouse model of the company.

A category of data warehouses for performance measurement can be distinguished, where the business process execution data is stored. The systems already mentioned use workflow data as one of data sources, but workflow data warehouse [7] represents the concept of Data Warehouse of Processes. The authors of Workflow Data Warehouse [7] argue why and when data warehouse can become an appropriate solution for storing and analyzing log files of process execution.

Methodologies how the performance should be evaluated also are a subject of research. For example, methodology [8], based on dynamic process performance evaluation, proposes measurement models for analysis of different process flows in order to control the quality of process execution. Activity flow, information flow, resource flow and others are measured using time, quality, service, cost, speed, efficiency, and importance as evaluation criteria.

Our approach uses the advantages of an existing data warehouse – ETL processes, analysis tools, data storage schemas – that allow to prepare and store indicators according to the different perspectives, as well as integrates the data warehouse with a performance measurement framework that is based on the life-cycle of indicators, which ensures the quality of the performance measurement by supplying necessary information for each measurement task.

2. Indicators and Their Life-Cycle

In our previous research [3] we defined a lifecycle of indicators, which consists of five steps – indicator definition, measurement, analysis, reaction and improvement. In each step an indicator is characterized by a different set of properties.

Indicator definition step describes the information needs of the user. In the measurement step the indicators get the values. The analysis step represents the process, when indicators are used to make decisions. The reaction step represents the implementation of decisions. The life-cycle ends with the evaluation of indicator definitions and predefined values of indicator properties during the improvement step.

2.1. Groups of Indicator Aspects

The properties of indicators are grouped in aspects according to the particular step (Figure 1). The explanation of the meaning of properties can be found in [3].

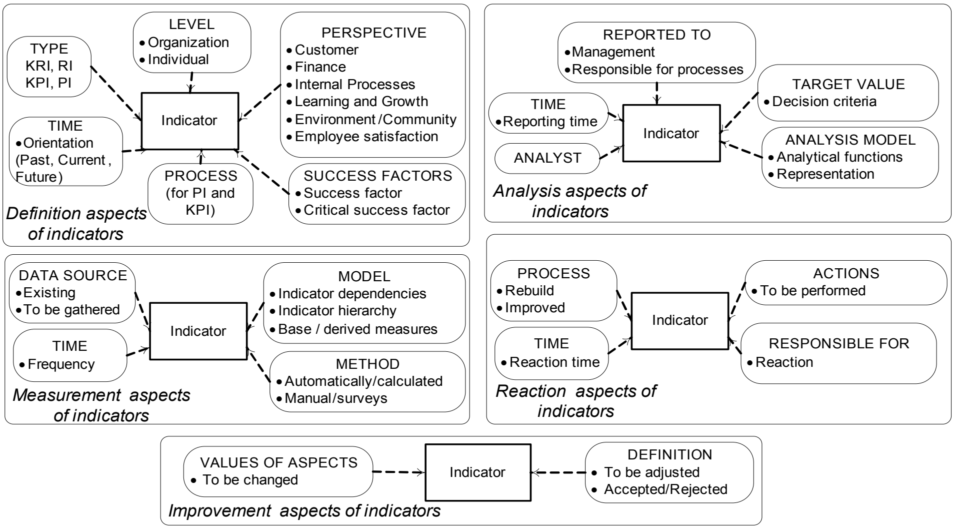


Figure 1. Five groups of indicator aspects [3]

One of the questions raised by the proposed measurement framework with indicators as the central element of interest was how the indicators should be formalized to bring the maximum of clarity into the measurement process – what, why and how is measured. Our previous research [3] is focused on the formal definition of indicators. The formalization method of sentences that expresses the indicators was proposed.

2.2. Formal Model for Indicator Definition

On one hand, indicators are the focus of data analysis in the measurement process. On the other hand, data warehousing models are built to represent the information needs for data analysis. Therefore we could talk about indicators as an information requirement for a data warehouse system.

The type of an information system to be developed has some impact on a way of formulating sentences that express requirements. We assumed that requirements for data warehouses and information requirements particularly have a similar structure or pattern. We based the proposed model on the structure evaluation of the sentences that formulate performance indicators taken from the performance measures database [1].

All indicators have common structure, for that reason it is possible to determine a pattern for re-writing business requirements formally. The requirement formalization may be represented as a metamodel. The detailed description of the metamodel and the algorithm how the sentences expressing the indicators are reformulated according to the given metamodel can be found in [3].

3. Reporting Tool and its Metamodel

One important and integral part of our process measurement framework is a reporting tool developed at the University of Latvia. This reporting tool is developed as the part of the data warehouse framework [9]. The reporting tool is based on metadata and in the latest version it has five metadata layers (Figure 2) that describe different aspects of defining and storage of a data warehouse schemata as well as defining and operating reports defined on these schemata.

Semantic, Logical and Physical metadata describe the data warehouse schemata in different levels of abstraction, starting from the business understanding of the schema elements, describing it by means of OLAP concepts at logical level and ending with the physical storage of the data warehouse tables. OLAP Preferences metadata is introduced to describe the user preferences on reports' structure and data and is used for OLAP personalization purposes. Reporting metadata contains definitions of reports on data warehouse schemata.

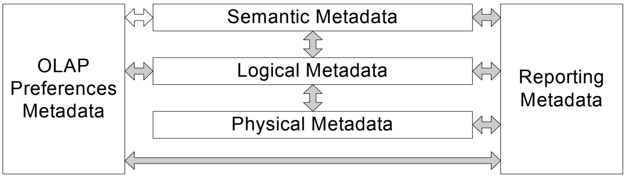


Figure 2. Metadata connections [10]

Metadata levels are interconnected by associations between particular classes of metadata. In the context of this research, the Logical and Reporting metadata are of particular interest; so both the levels as well as the connections between them will be described in more detail here. Detailed description of the rest of the metadata levels can be found in our previous research [11, 12].

3.1. Logical Level Metadata

The logical level metadata describes the data warehouse schema from the multidimensional viewpoint (Figure 3) and mostly is based on the OLAP package of Common Warehouse Metamodel (CWM) [13]. Therefore it contains the core concepts of OLAP – dimensions and fact tables (cubes in CWM).

Fact tables contain measures, but dimensions consist of attributes and hierarchies built from hierarchy levels. Fact tables and dimensions have *FactTableDimension association*. Only dimensions and fact tables having *FactTableDimension* associations can be used simultaneously in one report. More about connections with reporting metadata is given in section 3.3.

The standard OLAP package of CWM is extended by the class *AcceptableAggregation* that allows only meaningful definitions of aggregate functions (e.g. SUM, AVG) for each measure and dimension. This metadata is used to ensure correct queries by the reporting tool.

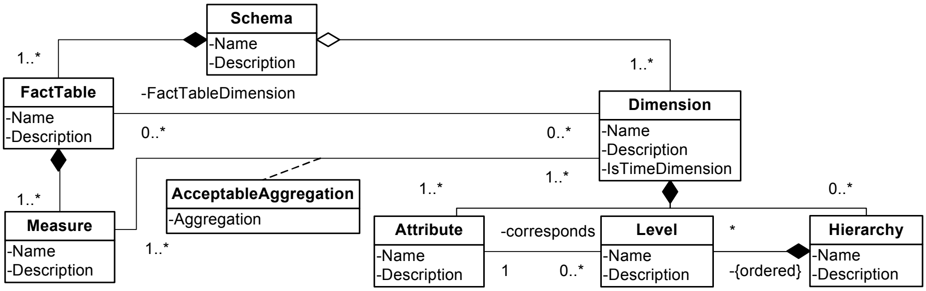


Figure 3. Logical level metadata [11]

3.2. Reporting Level Metadata

Reporting metadata describes the structure of reports (Figure 4). In the meaning of this model, reports are worksheets. Worksheets contain data items defined by calculations. Calculations in their turn specify formulas containing parameters and table columns that correspond to schema elements of the underlying data warehouse. Reports also are based on joins between tables and may have user-defined conditions.

Reports in the tool are created by choosing desired elements of a data warehouse schema and defining conditions, parameters etc. Only measures and attributes belonging to one schema could be included in the definition of one report.

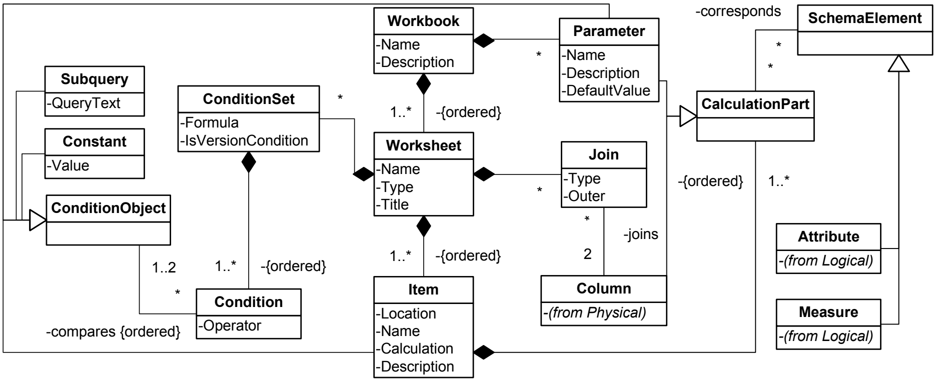


Figure 4. Reporting metadata [14, 12]

3.3. Connections between Logical and Reporting Metadata

The models of logical and reporting metadata are interrelated. Report items are defined by formulas from calculation parts. If a calculation part corresponds to a particular dimension attribute or measure, then this schema element from Logical metadata is connected to the class *CalculationPart* by the association *'corresponds'* in the reporting metadata.

4. Construction of Performance Measurement Framework

We propose an approach of building Performance measurement systems by substantially exploiting existing data warehouse technologies.

The proposed Performance measurement framework is grounded on the following principles of design and operation:

- processing of indicator information is performed in conformance with the life-cycle of indicators and formal indicator metamodel, defined in [3];
- measurement data are obtained through an ETL process and stored in a data warehouse;
- indicator analysis aspect is provided by using a ready-made data warehouse reporting tool extensively both for obtaining actual value from measurement data and providing users with detailed reports.

4.1. Architecture of Performance Measurement System

The kernel of the performance measurement framework (Figure 5) consists of performance management component and the indicator life-cycle support database, as well as the dashboard module.

Indicator life-cycle support database (detailed information is given in the next section) stores links to the formal definitions of indicators from the *Indicator formal definition database*, which is built according to the formal model for indicator definition described earlier in section 2.2. These indicator definitions are collected and formalized during the requirements gathering process for obtaining the precise and appropriate indicators for process measurement, as well as for documenting the information requirements of a data warehouse.

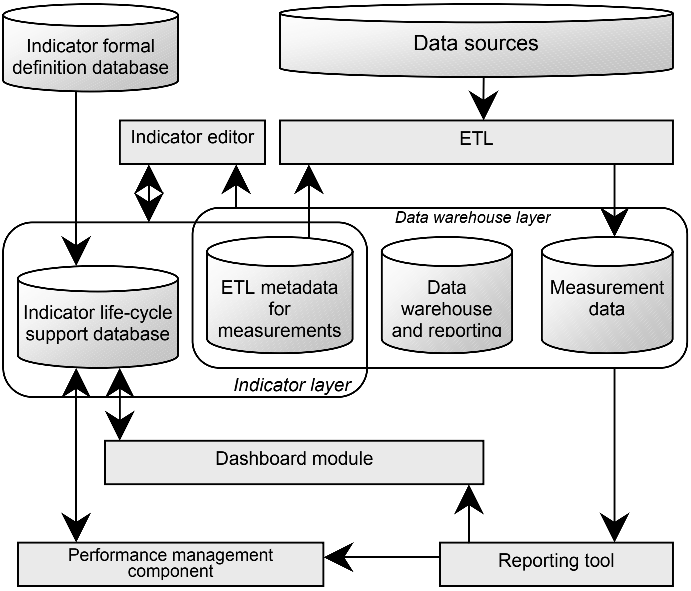

Indicator formal definition database
Data sources
Indicator editor
ETL
Data warehouse layer
ETL metadata for measurements
Data warehouse and reporting
Measurement data
Indicator life-cycle support database
Indicator layer
Dashboard module
Performance management component
Reporting tool

Figure 5. Data flow diagram of the performance measurement framework

Indicator editor is an administrative tool and is meant for two purposes: (1) to establish the links between the Indicator layer of the system and the Indicator formal definition database and (2) to configure the Indicator measurement (or ETL) metadata.

The measurement process, during which indicators get their values, is performed through the *ETL processes,* which use corresponding *ETL metadata for measurements*. The ETL component is an external part of the performance measurement framework. In the context of this research we assume that a set of procedures is defined for performing the data warehouse data renewal according to the values of ETL metadata for measurements (e.g. according to the planned timing schedule). During the ETL processes data from external *Data sources* are processed and loaded into the data warehouse that represents in our framework the *Measurement data*.

The remaining part of the data warehouse layer of the proposed framework is the *Data warehouse and reporting metadata* component that is developed according the previously described metadata layers in sections 2.1 and 2.2 that describe respectively the logical level of data warehouse schema and the reporting metadata.

Performance management component is the main part of the framework that is provided to coordinate the monitoring of business processes by analyzing the measurement results of indicators. Component is based on descriptions of different properties of indicators that are stored in the Indicator life-cycle support database and that allow the user to analyze the indicator values in the most appropriate way by means of two other components of the framework – *Dashboard module* and the *Reporting tool*.

The Dashboard module visualizes the most important values of indicators comparing them to the stored target values of indicators. The reporting tool provides more detailed information to the user by calling predefined reports linked to particular indicator definition. An existing reporting tool is used, which is built according to the previous mentioned reporting metadata (more information about this tool can be found in [14, 12].

4.2. Indicator Life-Cycle Support Database

Indicator life-cycle support database spins around the '*Indicator life-cycle support metadata'* (Figure 6), which define the behavior of the framework. These metadata are used by *Performance management component* designed to coordinate the workflow of the indicator life-cycle.

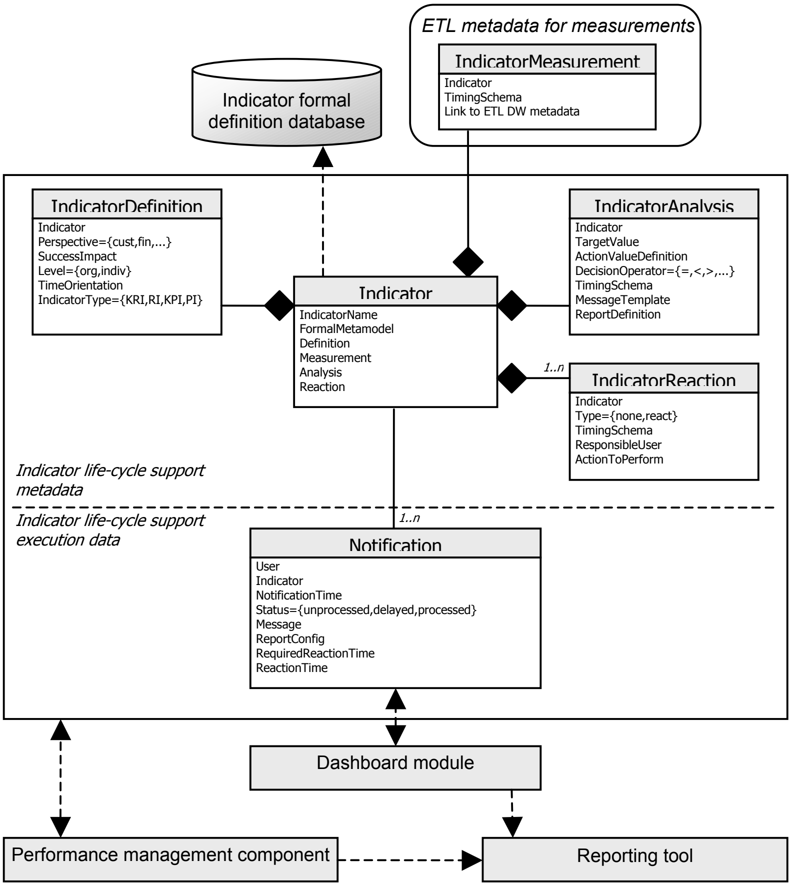


Figure 6. Indicator life-cycle support database and the context

As the duty of performing measurements is fully assigned to the data warehouse, ETL metadata are prepared and stored separately from Indicator life-cycle support database. Actually, this is one of the key points of the framework to fully connect data warehouse for such functionality.

Workflow status is stored in the 'Indicator life-cycle support execution data' and is accessible directly by users via Dashboard module. Workflow status is controlled both by Performance management component and by Dashboard module. It incorporates information about notifications by the system sent to users and the reaction of users to them.

4.3. Indicator Life-Cycle Steps in Performance Measurement Framework

According to the indicator life-cycle definition in section 2, Performance measurement framework should support all five steps of the life-cycle. This section is to describe the proposed framework according to the life-cycle steps.

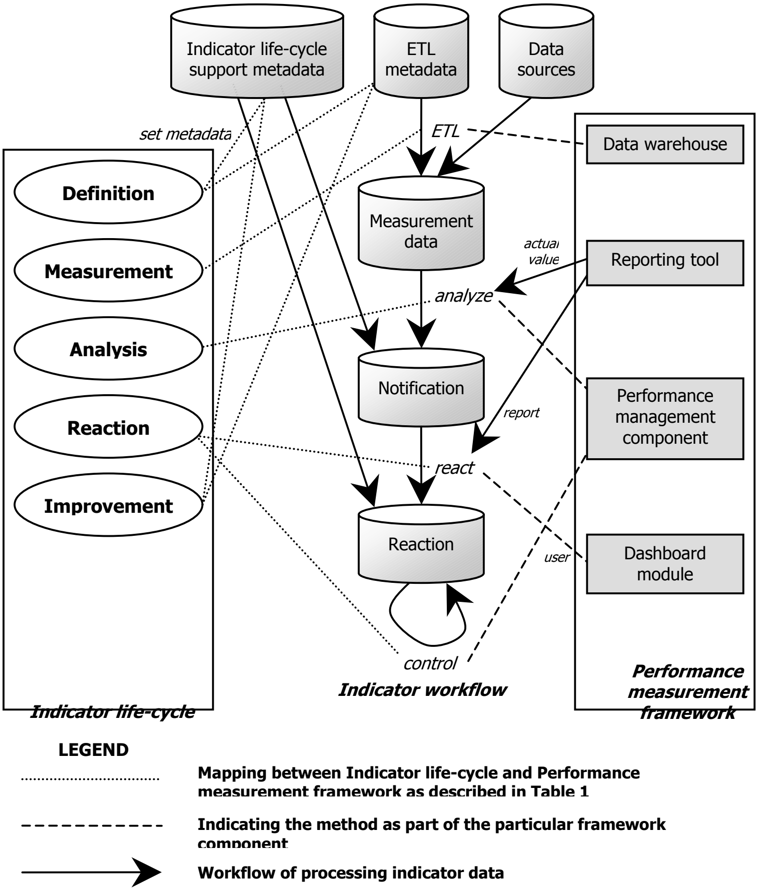

Indicator life-cycle support metadata
ETL metadata
Data sources
set metadata
ETL
Data warehouse
Definition
Measurement
Analysis
Reaction
Improvement
Measurement data
actual value
Reporting tool
analyze
Notification
report
Performance management component
react
Reaction
user
Dashboard module
control
Indicator life-cycle
Indicator workflow
Performance measurement framework
LEGEND
Mapping between Indicator life-cycle and Performance measurement framework as described in Table 1
Indicating the method as part of the particular framework component
Workflow of processing indicator data

Figure 7. Indicator life-cycle support implemented by the Performance measurement framework

Figure 7 shows the connections between each step and the workflow performed by the framework components and different data. The workflow of processing indicator data is organized in the following steps:

1. Measurement step is performed by an ETL process of the data warehouse (see section 4.4).
2. In analysis step measurement data are processed according to indicator life-cycle support metadata by the Performance management component (see the algorithm in Figure 8). Reporting tool is used here to obtain the actual value of the indicator. During this step, a record is added to the indicator life-cycle execution data; thus, the information about the performed measurements of indicators in form of a notification becomes visible to appropriate users in a special dashboard.

3. User's reaction is obtained from the Dashboard module (Figure 9) and can be of two types:
    - A request for the detailed notification. Reporting tool is used here to obtain a report that describes the actual measurement in detail;
    - Reaction. If the description of an indicator provides for a response to the notification, user is required to assert this in time and in a special way.
4. In control step Performance management component checks whether users have responded to the notifications, if such reactions were appointed in the analysis step (Figure 10).

The above described processing of indicator data by Performance measurement framework is performed in conformance with the life-cycle of indicators.

Table 1 shows mapping between the indicator life-cycle and its implementation by the Performance measurement framework.

Table 1. Mapping "Indicator life-cycle ↔ Performance measurement framework"

Indicator life-cycle aspect group	*Description of implementation by the performance measurement framework*
1. Definition	Indicator definition is described in indicator formal definition database, as well as Indicator life-cycle support database. Indicator definition includes preparation of the metadata required to ensure the whole process.
2. Measurement	Measurement process is fully delegated to the data warehouse and its appropriate ETL process.
3. Analysis	Analysis is coordinated by Performance management component. Measurement data are processed and displayed to the users.
4. Reaction	A user reads and, if required, reacts to the notification. Performance management component controls the reaction.
5. Improvement	Indicator improvement technically matches indicator definition.

```
Procedure analyze
Begin
    Repeat Forever
        Foreach indicator From Indicator Do
            analysis := Indicator.Analysis
            Wait for the next report according to analysis.TimingSchema
            actual_value := run report according to analysis.ActualValueDefinition
            If analysis.DecisionOperator (actual_value, analysis.TargetValue) = True Then
                Forall reaction In Indicator.Reaction Do
                    add record to Notification With
                        User := reaction.ResponsibleUser
                        Indicator := reaction.ResponsibleUser
                        User := indicator
                        NotificationTime := current time
                        Status := 'unprocessed'
                        Message := compute according to analysis.MessageTemplate
                          and indicator.Definition and reaction.ActionToPerform
                        ReportConfig := set according to analysis.ReportDefinition
                        RequiredReactionTime := compute according to reaction.TimingSchema
                        ReadTime := Null
                        ReactionTime := Null
```

Figure 8. Algorithm of the analysis step in the Performance management component

```
Procedure react
Begin
    Foreach user
        Display all from Notification in the dashboard Where User = user
        Foreach notification From Notification Where User = user Do
            Wait for user action Do
                Case user asks to show detalized information Do
                    run report according to notification.ReportConfig and display it
                Case performs an action according to report.Indicator.Reaction.ActionToPerform Do
                    notification.ReactionTime = current time
```

Figure 9. Algorithm of the reaction step in the Dashboard module

```
Procedure control
Begin
    Repeat Forever
        Foreach notification From Notification Where ReportStatus <> 'processed' Do
            If notification.Indicator.Reaction.Type = 'none' Then
                notification.Status = 'processed'
            Else
                If notification.ReactionTime Is Not Null Then
                    notification.Status = 'processed'
                Else If current time > notification.RequiredReactionTime Then
                    notification.Status = 'delayed'
```

Figure 10. Algorithm of the control step in the Performance management component

4.4. Integration with Data Warehouse Components

ETL metadata for measurements (*IndicatorMeasurement* class) is a part of Indicator life-cycle support database (Figure 6). The *Indicator* attribute identifies a particular attribute that is measured, whereas *TimingSchema* describes the time parameters of measurement (e.g. frequency, exact starting time). The last attribute – *ETLprocess* – points to the data warehouse meatadata repository, particulary to the ETL metadata part of the repository that describes mappings between the source and data warehouse schemas. This metadata also contains calls to corresponding procedures that implement these mappings and necessary data transformations. For the proposed measurement framework we can assume that the *IndicatorMeasurement* class contains the procedure call that renews the data warehouse data schema that contains data necessary for calculation of the given indicator.

The *IndicatorAnalysis* class of Indicator life-cycle support database (Figure 6) and its *ReportDefinition* attribute is planned to be a pointer to the report definition stored in accordance to the metamodel of the reporting tool.

Reporting metadata (Figure 4) contain the *Worksheet* class that identifies a particular report that can be invoked when analysis of measurement results is performed. The report can be simple, when one particular value is retrieved to compare it with a target value, or complex, when the report is used for the detailed analysis. The complexity of the report depends on the definition of the particular report.

5. Conclusions

Using data warehouses in performance measurement systems has been already extensively explored. The proposed Performance measurement framework has been designed to obtain the maximum benefits from matured data warehouses technologies in implementing indicator life-cycle support.

The applied model of indicator life-cycle serves as a theoretical means of quality assurance for the performance measurement. The use of data warehouses as integral part of the framework covers two important aspects of ensuring the indicator life-cycle: (a) indicator measurement, and (b) part of indicator analysis (performed by Reporting module).

The provided method for performance measurement ensures timely and to given context appropriate decision making process. The indicator life-cycle support database stores metadata that define and schedule the measurement and control processes of indicators, including timing schemas, responsibilities and actions to be performed. The proposed framework provides the option to build performance control on the activities initiated from the side of the measurement system, as soon as the system recognizes the problem and so the need for more detailed analysis.

Preliminary works of implementing the framework are already in progress, so we expect the first experimental results in the near future.

Acknowledgments

This work has been supported by European Social Fund (ESF) project No. 2009/0216/1DP/1.1.1.2.0/09/APIA/VIAA/044.

References

[1] D. Parmenter, *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs*. Second ed. Jon Wiley & Sons, 2010.

[2] R. S. Kaplan and D. P. Norton, *The Balanced Scorecard*. Harvard Business School Press, 1996.

[3] A. Niedritis, L. Niedrite, and N. Kozmina, Performance measurement framework with formal indicator definitions. In: J. Grabis, M. Kirikova, editors, *Perspectives in Business Informatics Research*, Lecture Notes in Business Information Processing **90** (2011), Springer, Berlin, 44-58.

[4] B. List and K. Machaczek, Towards a corporate performance measurement system. In: *Proceedings of the ACM Symposium SAC'04*, ACM Press, 2004, 1344-1350.

[5] M. Jarke, T. List, and J. Koller, The challenge of process data warehousing. In: *Proceedings of the 26th International Conference VLDB'2000*, 2000, 473-483

[6] P. Kueng, T. Wettstein, and B. List, A holistic process performance analysis through a process data warehouse. In: *Proceedings of the 7th American Conference on Information Systems (AMCIS'01)*, 2001, 349-356.

[7] A. Bonifati, F. Casati, U. Dayal, and M. C. Shan, Warehousing workflow data: challenges and opportunities. In: *Proceedings of the 27th International Conference VLDB'2001*, 2001, 649-652

[8] W. Tan, W. Shen, L. Xu, B. Zhou, and L. Li, A business process intelligence system for enterprise process performance management, *IEEE Transactions on Systems, Man, and Cybernetics* **38**(6) (2008), 745-756.

[9] D. Solodovnikova, Data warehouse evolution framework. In: *Proceedings of the Spring Young Researcher's Colloquium On Database and Information Systems (SYRCoDIS'07)*, Moscow, Russia, 2007. Available from: http://ceur-ws.org/Vol-256/submission_4.pdf.

[10] N. Kozmina and D. Solodovņikova, Determining preferences from semantic metadata in OLAP reporting tool. In: L. Niedrite, R. Strazdina, B. Wangler, editors, *Perspectives in Business Informatics Research*, Local Proceedings of the 10th International Conference BIR 2011, Associated Workshops and Doctoral Consortium, Riga Technical University, 2011, 363-370.

[11] D. Solodovnikova, Metadata to support data warehouse evolution. In: *Proceedings of the 17th International Conference on Information Systems Development (ISD'08)*, Paphos, Cyprus, 2008, 627-635.

[12] D. Solodovnikova and L. Niedrite, Evolution-oriented user-centric data warehouse. In: J. Pokorny, V. Repa, and K. Richta, W. Wojtkowski, H. Linger, Ch. Barry, M. Lang, editors, *Proceedings of the 19th International Conference on Information Systems Development*, Springer, 2011, 721-734.

[13] *Common Warehouse Metamodel Specification, v1.1*. Object Management Group. Available from: http://www.omg.org/cgi-bin/doc?formal/03-03-02.

[14] D. Solodovnikova, Building Queries on Multiple Versions of Data Warehouse. In: H.-M. Haav, A. Kalja, editors, *Databases and Information Systems V*, Selected Papers from the 8th International Baltic Conference DB&IS 2008, IOS Press, 2008, 75-86.

Multi-Layered Architecture of Decision Support System for Monitoring of Dangerous Good Transportation

Dale DZEMYDIENE[a1] and Ramunas DZINDZALIETA[b]
[a]*Mykolas Romeris University, Ateities 20, LT-08303, Vilnius, Lithuania*
[b]*Institute of Mathematics and Informatics, Vilnius University, Akademijos 4, Vilnius, Lithuania*

Abstract. The consideration of this study is attached to the representation of knowledge content of dynamic application domain of transportation related to the risk evaluation of possible abnormal situations of dangerous good transportation. Multi-layered conceptual architecture is assembled by the models of knowledge representation at higher level including conceptual models of information structures, dynamic process analysis, and problem solving tasks in transportation processes of dangerous goods. The model represents behavioral analysis of target system based on Petri nets. The paper presents the technological platform how aggregate sensor components integrated with mobile technology can support the on-line processing of real data for localization and monitoring of transport objects and allow on-line recognition of abnormal situations. The representational platform describes a general component model that is a basis for expressing properties of knowledge of domain for informational structure specification.

Keywords. Decision support system (DSS), mobile control system, mobile sensor, intermodal transportation

Introduction

Over the past few years, a large number of researches had emerged approaching the use of mobile and other information management technologies in intermodal transportation area. While there is much literature about the logistic chain analysis [1, 5] and the intermodal transportation management itself [2, 3, 4] comparatively little has been written on this subject in relation to mobile sensor based technology implementation in intermodal container transportation.

Lee and Chan (2009) proposed a RFID-based reverse logistics framework and introduces genetic algorithm to optimize the locations of collection points for product returns in order to maximize the coverage of customers, which allow economically and ecologically reasonable recycling [8]. Jedermann *et al.* (2006) analyzed new sensor, communication and software technologies which were used to broaden the facilities of tracking and tracing systems for food transports, where an embedded assessing unit detected from sensor data collected by a wireless network potential risks for the freight quality and estimated the current maturing state of agricultural products which were supported by measurements of the gaseous hormone ethylene as an indicator for the ripening processes [9].

This research is inspired by e-Safety Initiative [22, 24]. We can find works related with hazardous materials transportation analysis [13, 14, 15, 16, 18, 19, 33]. Related

[1] Corresponding Author: E-mail: daledz@mruni.eu

works evaluate sensor readability [6] and integrate UHF RF powered chips by using sensors for wireless monitoring [5]. Our proposed multi-layered conceptual architecture assembles the technological platform of aggregating sensor components which are working by mobile technology and supports the on-line processing of real data for localization and monitoring of transport objects. The models of knowledge representation for the on-line recognition of abnormal situations are included in the decision support system (DSS) under development by conceptual models of information structures, dynamic process description and decision making tasks.

The representational platform describes a general component model that is a basis for expressing properties of knowledge by using Petri nets which consideration aimed at helping in management processes of multi-modal transportation. The analysis of transportation processes as complex technology has been proposed by means of imitation modeling [11, 25]. The attention is paid to the representation of dynamic and static aspects of a target system. The approach of using integrated conceptual models such as semantic models for representing information structures and fuzzy logic Petri nets [10] for functional analysis is focused on the consideration of temporal aspects of domain. However, multimodal transportation conditions, information security and other risk issues are less analyzed making them the primary objectives of the proposed transportation management mobile control system. In this paper, we study an emerging field of intermodal transportation and offer a combined RFID and mobile sensor based mobile control system, to ensure seamless end-to-end tracking and visibility from global to local level in intermodal transportation management by evaluating the potential risks involved in transportation of dangerous goods.

The aims of this research concern the construction of knowledge base for risk description of transportation dangerous goods and relation it with decision making deriving actions according to the data from sensors working on-line as the monitoring subsystem of transport objects. The tasks of this research are:

- to choice the knowledge representation techniques for description of risk of transportation using recognition mechanism using information about transport mean mobility and sensor parameters;
- to integrate the risk management component into the decision support of transportation processes;
- to present the architecture of the decision support system working as monitoring on-line system using mobile technologies;
- for assuring a high level of information and transportation security, and improving the efficiency of the communication to upgrade the capability of the general information system by integrating the mobile interaction system, using SIP (Session Initialization Protocol).

For the construction of knowledge base we are choosing Petri nets for the description of imitational model of transportation system. Petri nets are used to describe decision making processes and SIP communication protocol. For the semantic representation of data we are used class diagrams based on object oriented model of UML. For risk representation and possibilities of evaluation the levels of risk we describe the set $S = \left\{ \, s_k \, \right\}$ of types of scenarios of accident events of transportation which we can to recognize. Scenarios are described using the probability of evolving of such type s_k of scenarios.

1. The Architecture of Decision Support System with Embedded Subsystems for Monitoring and Localization of Moving Objects

The main design principles of DSS framework is presented by means of conceptually layered framework, with a view to associate the functionality of implemented components of the subsystems in the existing framework of the DSS (Figure 1). The system is dividing into various layers. We depict different context models used for representing, storing and exchanging sensing for contextual information representation needed for support decisions.

The real-time working subsystems (monitoring of data part) are embedded in the target system as a concurrent computing system related with the monitoring of sensors (Figure 1).

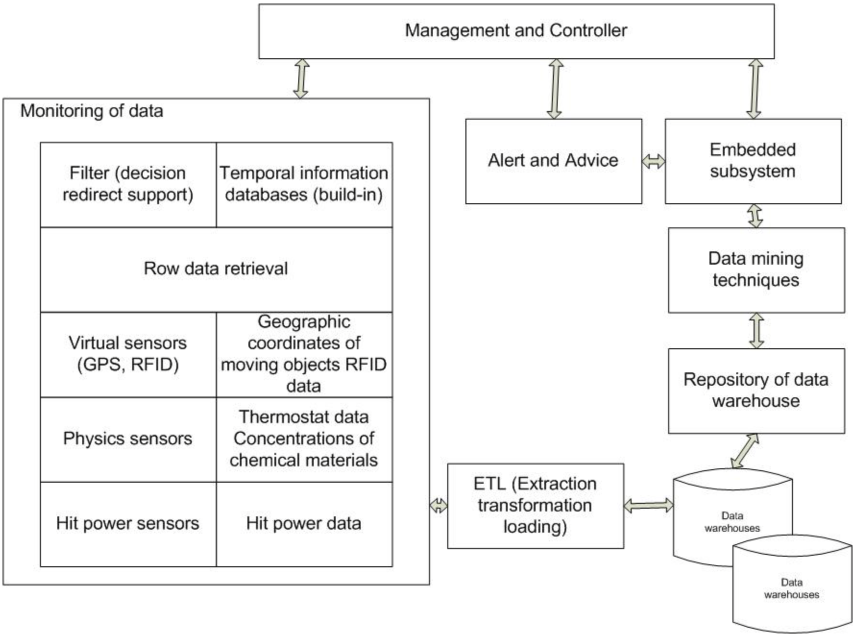

Management and Controller
Monitoring of data
Filter (decision redirect support)
Temporal information databases (build-in)
Row data retrieval
Virtual sensors (GPS, RFID)
Geographic coordinates of moving objects RFID data
Physics sensors
Thermostat data Concentrations of chemical materials
Hit power sensors
Hit power data
Alert and Advice
Embedded subsystem
Data mining techniques
Repository of data warehouse
ETL (Extraction transformation loading)
Data warehouses
Data warehouses

Figure 1. Architecture of main components of the DSS

The monitoring subsystem connected with the expert subsystem must detect the faults of process performance. The time for obtaining a solution is often strictly limited. These conditions impose strict deadlines on the obtaining a decision and maintaining the functioning correctness. The system behavior defines a set of temporal dependencies, dynamic evaluation of situations, adaptively control feedbacks and complexity management, which must be implemented in the embedded DSS, according to related works [11, 12, 21]. The system works as a multiple agent based system.

The monitoring component of the system integrates several sensor systems which observe the transportation means and indicate possible conditions of the state. Such sensors are aimed at localization of the object, observing the main physical parameters inside the object, which can characterize multiplex state evaluation. The main types of sensors are represented in Figure 1. Such data became row data for transforming them the data warehouses. The metadata represented in the conceptual schema of the repository of data warehouses are introduced in our system for a better understanding of data semantics and contextual information. The extraction transformation loading (ETL) engine is used for revealing and storing such row data into the data warehouses. Data mining techniques are introduced in the DSS as the components for extraction of the main rules and patterns of the situation recognition which can help to integrate multi-dimensional parameters into decision support processes and control processes of the accident event situation.

2. Description of Functional Requirements and Risk of Transportation System

The integration possibilities of knowledge representation techniques: semantic model constructions, macro Petri nets with imitational interpretations of processes are considered. Such common interaction model is illustrated in Figure 2. The whole modelling system of multi-modal transportation was divided into three subsystems: ”Environment”, “Node” and “Chain”. The structural scheme of component application for multi-modal transportation system is presented in Figure 2.

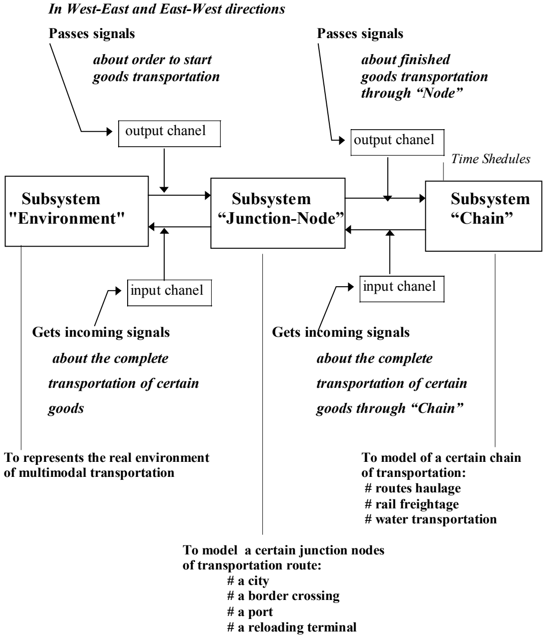

In West-East and East-West directions
Passes signals
about order to start goods transportation
Passes signals
about finished goods transportation through “Node”
output chanel
output chanel
Time Shedules
Subsystem "Environment"
Subsystem “Junction-Node”
Subsystem “Chain”
input chanel
input chanel
Gets incoming signals
about the complete transportation of certain goods
Gets incoming signals
about the complete transportation of certain goods through “Chain”
To represents the real environment of multimodal transportation
To model of a certain chain of transportation:
# routes haulage
# rail freightage
# water transportation
To model a certain junction nodes of transportation route:
# a city
# a border crossing
# a port
# a reloading terminal

Figure 2. The structural scheme of interaction of aggregate components for intermodal transportation representation

Decision making is performed considering a lot of various factors: evaluating technical infrastructure of multi-modal transportation and organizational aspects, comparing reports with the real situation.

The subsystem "Environment" is dedicated to model the real environment of multi-modal transportation. It has two output channels with the first and the last junction-nodes of the "Node" subsystem, through which the "Environment" passes the output signal about the order to start goods transportation accordingly to the West-East or East-West direction.

The route is divided into road stretches and each is characterized by different characteristics. Risk is related with scenarios of accident events, influenced by types of dangerous goods, and surroundings. The approaches of multiple complex description of scenarios influence the classification them by types and can be based on the ontology of this phenomenon. Federal and provincial legislation provide for the regulation of an extensive list of products, substances or organisms classified as dangerous. The products fall into one of nine classes: explosive, flammable, radioactive, etc. The model is focused on evaluation of a proper frequency of accidents, following [15, 16].

The set $S = \{ s_k \}$ represents types of scenarios of accident events of transportation which we can to recognize, where $k = \overline{1,n}$.

Following the recommendations of approaches by [15, 17, 24], the expected number of fatalities as a consequence of an accident occurred on the road stretch *r* and evolving according to a scenario s*ₖ*, can be expressed as:

$$\mathrm{B}_r = \sum_{k=1}^{n} f_r N_{r,s_k} P(s_k) \qquad (1)$$

where f_r is the frequency of accident in the *r*-th road stretch [accident·year^{-1}], N_{r,s_k} is the number of fatalities according to a scenario s_k in the *r*-th road stretch [accident fatalities^{-1}], $P(s_k)$ is the probability of evolving scenarios of type s_k, following the accident (i.e. collision; roll-over; failure, etc.).

The transportation network can be considered as a number of junctions (nodes) linked one to another by a number of arcs (Figure 3).

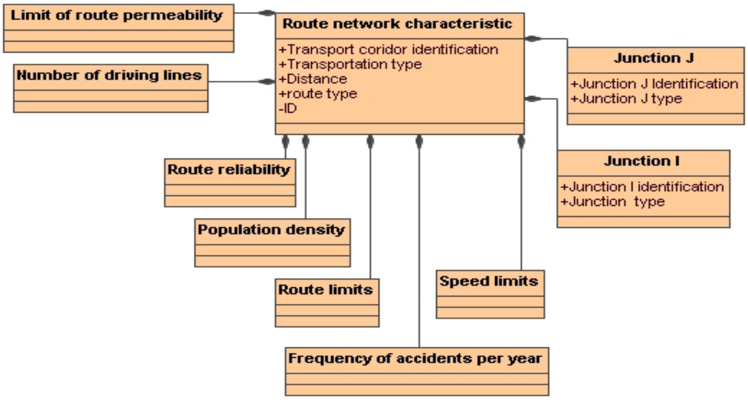


Figure 3. The example of class diagram for conceptual representation of main parameters of route stretch characteristics

The junctions represent the cross roads, towns, tool-gates, storage areas, etc. in the transportation network. An arc between two junctions can be characterized by a different number of road stretches and the expected number of fatalities for the arc is:

$$B = \sum_{r} \sum_{s_k} f_r N_{r,s_k} P(s_k) \qquad (2)$$

The frequency of an accident involving the scenario s_k, on the r-th road stretch, can be expressed as:

$$f_{r,s_k} = f_r P(s_k) \tag{3}$$

$$f_r = \gamma_r L_r n_r \text{ , where:} \tag{4}$$

$$\gamma_r = \gamma_{0,r} G \text{,} \tag{5}$$

where γ_r is the expected frequency on the r-th road stretch [accident·km^{-1}·vehicle^{-1}·year^{-1}], L_r is the road length [km], n_r is the number of vehicles through the road r-th stretch in [vehicle], $\gamma_{0,r}$ is the regional accident frequency [accident·km^{-1}·vehicle^{-1}·year^{-1}], according to [17].

G is probabilistic parameter, characterized as a common evaluation parameter of environment. Various factors influence the accident events: environmental, behavioral, physical, mechanical, Road intrinsic descriptors are described by these parameters.

$$G = \prod_{j=1}^{m} G_j \tag{6}$$

where G is the local enhancing/mitigating parameter. The main types of these parameters we can describe as: G_1 is a parameter depending on temperature, G_2 is a parameter that depends on the inherent factor (such as tunnel, bend radii, slope, height gradient, etc), G_3 is a parameter that depends on the metrological factor (such as snow, sun, rain, ice, etc), G_4 is a parameter that depends on the wind speed and wind direction, and others until such parameter that we can recognize G_m **.**

N_{r,s_k} is the total number of fatalities according to Eq. (2):

$$N_{r,s_k} = (\Phi^{in}_{s_k} o^{\Delta t} v_r + \Phi^{off}_{s_k} d_r) P(F, s_k) \tag{7}$$

Being the in-road and the off-road number of fatalities calculated, respectively, as:

$$N^{in}{}_{r,s_k} = \Phi^{in}_{s_k} o^{\Delta t} v_r P(F, s_k) \tag{8}$$

$$N^{off}{}_{r,s_k} = \Phi^{off}_{s_k} d_r P(F, s_k) \tag{9}$$

where $\Phi^{in}_{s_k}$ is a consequence of the in-road area associated with scenario s_k [m^2];

$\Phi^{off}_{s_k}$ is a consequence of the off-road area associated with scenario s_k [km^2];

P(F,s$_k$) is a probability of fatality *F* for accident scenario s_k ;

$o^{\Delta t}$ is the average vehicle occupation factor during specific time period Δt, which can depend on the seasons or day time;

v_r is the vehicle density on the road area [vehicle·m^{-2}];

d_r is the population density of the *r-th* road area environment [inhabitants·km^2].

Also the "Environment" has two incoming channels with the first and the last junction nodes of "Junction-Node" subsystem, through which the "Environment" gets the incoming signal about the complete transportation of the certain goods. The subsystem "Junction-Node" is dedicated to model the certain node of the transportation route, i.e. a port, a city or a border crossing. Each node from such subsystem has out-coming channels with "Chain" subsystem, through which the " Junction-Node" passes the out-coming signal about the finished goods transportation through the " Junction-Node" (goods loading, warehousing, customs).

The results of analysis of attractiveness of the transport system between forwarding agents showed that the most important evaluation criteria are: transport cost, reliability and lead time of transportation. The weight of these three factors is varying among different respondents. It is linked to the nature of cargo being carried and depends on special requirements of senders and so on. Also the basic cargo compatibility characteristics must be taken into account while allocating cargo in a container because the interrelationships between the transport properties of cargos may result in quality degradation and damage. Different cargo may react with one another and possibly with their environment. Most changes of the cargo occurring during transport are unwanted and considered damage. Cargo properties are described by their characteristic features, specific functions, utility value and its quality and etc., where transport properties cover the properties of a cargo which need to be taken into account to ensure value loss-free intermodal container transportation.

The evaluation and selection of route also depends on the type of loads and on the desirable duration of transportation. Information accumulated in the system should help to determine technical state and reliability of routes, transportation duration. In order to select the optimal route of transportation the price of transportation and reliability of the route play an important role as well.

Reliability of the route is a complex evaluation and it is not easily determined. It should reflect assurance of load safety, possibility of assault, assurance of freight delivery in the limits of fixed terms.

3. Representation of Transportation Process Imitational Model using Petri Nets

The computing results of reasoning were obtained by application of logical Petri nets [10]. Classical Petri nets are defined as a structure $N = <S, T, F>$ where S means set of places, T is set of transitions and F is function of transition works.

$F \subseteq (S \, x \, T) \cup (T \, x \, S)$, where $(\forall \, t \in T) \, (\exists \; p, q \in S)(p, t), (t, q) \in F.$

Graphical representation of Petri nets is set up by the following symbols: *places* - by rings, *transitions* - by rectangles, and *relations* – by pointers between transitions and places or places and transitions. In classical Petri nets, there is a token placed if the expression is true (1) or not if it is false (0).

Let $FLPN=(P,T,F,M_0,D,h,\alpha,\theta,\lambda)$ be a fuzzy logical Petri net. Set of places $P_0 = \{p| \, M_0 \, (p)>0 \wedge \forall p \in P\}$ is called a set of places of initial true propositions. D_0 corresponding with P$_0$ is called a set of initial true propositions. Function h_s: $P_s \rightarrow D_s$ is an association function, representing a bijective mapping from places to propositions. Propositions, such that $h_s(p)= h_s(p), \; \forall p \in P_s$. Function α_s: $P_s \rightarrow [0,1]$ is an association function, representing a mapping from places to real values between 0 and 1, such that $\alpha_s(p)= \alpha$ $_s(p)$, $\forall p \in P_s$; $\theta_s,\lambda_s : \mathrm{T}_s \rightarrow [0,1]$ are association functions, representing a mapping from transition to real values between 0 and 1, such that $\theta_s (t) = \theta (t)$, $\lambda_s(t) = \lambda(t)$, $\forall t \in T_s$. The firing rules are the same as in classical Petri nets.

The exceptional feature is the fact that the net transition can represent a sequence of smaller operations with transition parameters connected with the processes. It is possible to consider the net as a relation on (E,M_0,Ξ,Q, Ψ), where E is a connected set of locations over a set of permissible transition schemes, E is denoted by a four-tupple: $E=(L,P,R,A)$, where L is a set of locations, P is the set of peripheral locations, R is a set of resolution locations, A is a finite, non-empty set of transition declarations; M_o is an initial marking of a net by tokens; $\Xi=\{\xi j\}$ is a set of token parameters; Q is a set of transition procedures; Ψ is a set of procedures of resolution locations.

The net transition is denoted as $a_i =(s_{i,} t(a_i),q_i)$, where $\mathrm{s_i}$ is a transition scheme, $t(a_i)$ is a transition time and q_i is a transition procedure. In order to represent the dynamic aspects of complex processes and their control in changing environment it is impossible to restrict ourselves on the using only one temporal parameter $t(a_i)$ which describes the delaying of the activity, i.e. the duration of transition. The input locations L_i' of the transition correspond to the pre-conditions of the activity, and the output locations L_i'' correspond to post-conditions of the activity. The complex rules of transition firing are specified in the procedures of resolution locations Ψ and express the rules of process determination.

Any *IF-THEN* rule is given of the form of:

$\mathrm{IF}\ X_1 \text{ is } A_1 \text{ AND} \ldots \text{AND } X_n \text{ is } A_n \text{ THEN } Y \text{ is } B$, where $A_1,.., A_n$ and B are certain predicates characterizing the variables $X_1,...,X_n$ and Y.

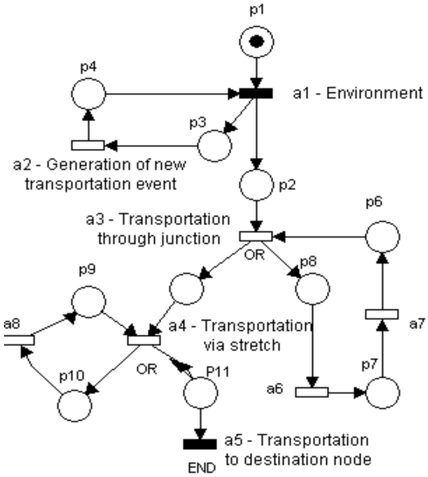


Figure 4. An example of description of transportation chain by means of macro-Petri nets

The set of *IF-THEN* rules forms linguistic description:

$$\begin{array}{l} R_1 := \mathrm{IF}\ X_1 \text{ is } A_{11} \text{ AND} \ldots \text{AND } X_n \text{ is } A_{1n} \text{ THEN } Y \text{ is } B_1 \\ \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ R_m := \mathrm{IF}\ X_1 \text{ is } A_{m1} \text{ AND} \ldots \text{AND } X_n \text{ is } A_{mn} \text{ THEN } Y \text{ is } B_m \end{array} \tag{10}$$

where each transition of the result of fuzzy Petri net corresponds to one rule of such linguistic description.

4. Integration of Localization and Sensing Information of the Transport Objects in DSS

Moving objects are constrained by a road network and they are capable to obtain their positions from an associated GPS receiver [28]. Moving objects (termed as mobile clients) are recognized by their location information. Location server and the central data warehouse are in the server site. The relationship is possible via a wireless communication network [11, 20]. The disconnection between client and server is realized by other mechanisms in the network than the tracking. The disconnection occurrences activate mechanisms which notify the server which appropriate actions are needed. After each update from a moving object, the position is represented in the data warehouse and the system informs the moving object about the location. The moving object issues an update when the predicted position deviates by some threshold from the real position obtained from the GPS receiver [26, 27, 28, 29].

The client initially obtains its location information from the GPS receiver and from the physical and virtual sensors [30, 31]. This possibility allows collection of the data from the sensors and processes them on-line. The data of sensor parameters are exchanged, and then the event e_{ti} influence changes in reality. If the data are changed critically, DSS gets a signal or message. The architecture of these components is represented in Figure 5 and 6.

To combine the web service protocol, e.g. simple object access protocol (SOAP), with SIP is very important for securing the communication between server systems and mobile devices [20, 23]. SOAP can be used on top of SIP or in parallel in the same layer. SIP is defined to be used only as a signaling protocol in the application layer. Thus, work is focused on the use of SIP on the control (signaling) plane in parallel of SOAP on the user plane according to [27].

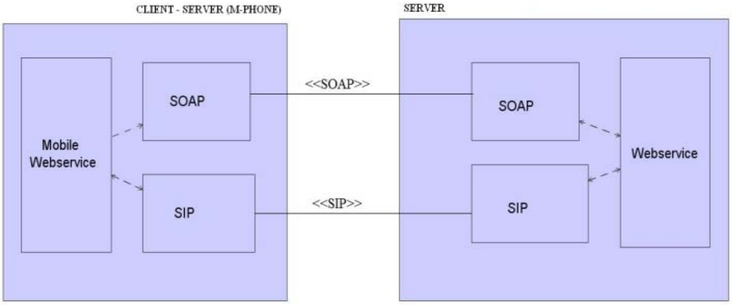


Figure 5. The scheme of integration of mobile Web services and a SIP user agent

Separation the user and signaling plane has advantages with respect to protocol design, communication software design, and performance. SIP is used to transmit “application layer” signaling messages.

In order to communicate between two different mobile devices via Web Service there must be a mobile web service endpoint. The mobile web service endpoint is a SIP URI (URI is based on the IP address). In generally, each terminal is able to provide and use mobile web services (MWS) at the same time and within the same SIP session. The use of MWS in a P2P manner is possible by establishing a SIP session between the devices. The MWS endpoints are SIP URIs, the web services endpoints of both clients are URIs containing the current IP address. First, we need a set of building-block of web services. They are common basic web services required by most mobile-service applications. The MWS and proxies have to register to the SIP agent in order to be notified about URI (IP address) changes (Figure 6).

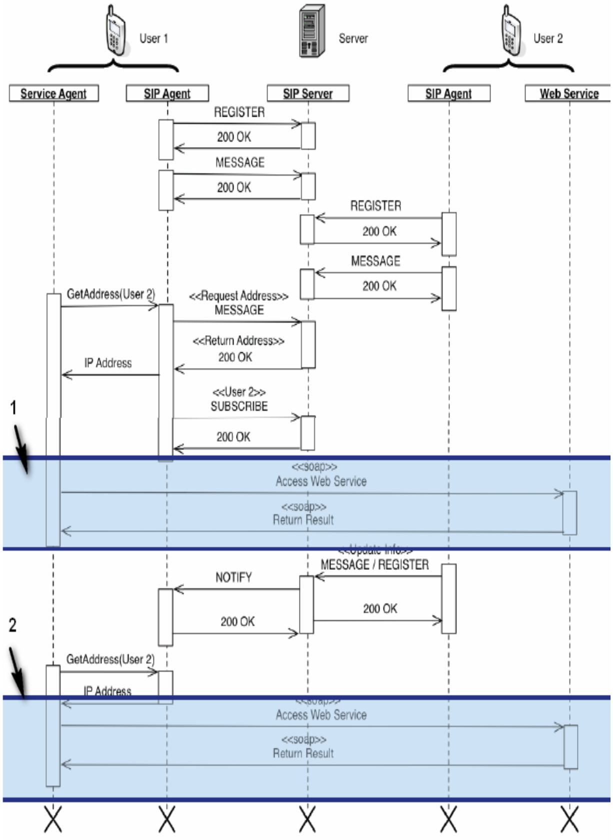


Figure 6. Sequence diagram of messaging of the connection session between SIP agents

The user of mobile device must share its physical address with the registrar in the network. Along with this "registration" is the public identity that is to be bound to the physical address (Figure 7).

The public URI can change physical addresses many times as a subscriber moves about the network, so the binding of addresses may change frequently. The connection of two participators is able to start by sending a SIP *INVITE* message after starting the SIP session between two devices (or conference). This session is initialized by request that enables a virtual connection between two or more entities for exchange of user data. Registration is not required for the agents using a proxy server for outgoing calls. It is necessary, however, for an agent to register the receipt of income calls from proxies.

The sensor's subsystems are worked as agents in parallel and the important information is writing on the temporal information registration window (TIRW). The process control subsystem of the DSS must detect such facts: what the maximum value was in concrete time interval in surroundings, the number of times a value exceeded a predefined reference value (i.e., the limitations of concentrations of harmful materials in the surrounding, sewerage water, etc.), the temporal delay between the maximum of a variable, and the maximum effect on another variable (Figure 9).

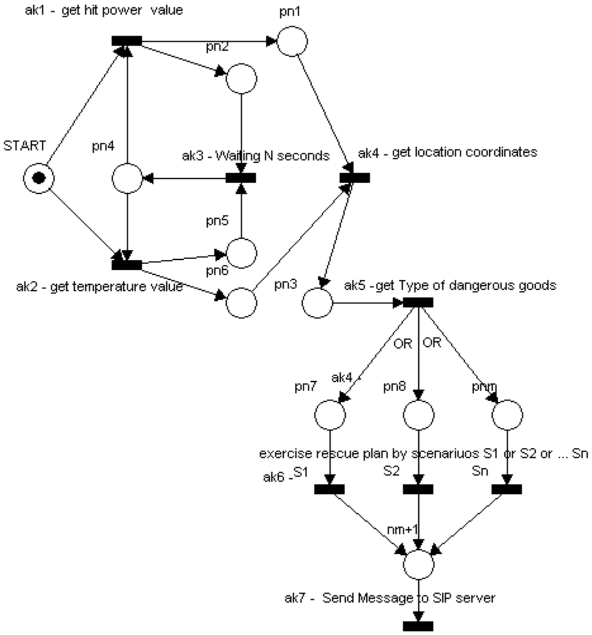

ak1 - get hit power value
pn1
pn2
START
pn4
ak3 - Waiting N seconds
ak4 - get location coordinates
pn5
pn6
ak2 - get temperature value
pn3
ak5 -get Type of dangerous goods
OR
OR
pn7
ak4
pn8
pnn
exercise rescue plan by scenariuos S1 or S2 or ... Sn
ak6 -S1
S2
Sn
nm+1
ak7 - Send Message to SIP server

Figure 7. The Petri nets schema of monitoring processes and connection with SIP

Active RFID tags are also constantly powered, whether in range of a reader or not, and are therefore able to continuously monitor and record mobile sensor status, particularly valuable in measuring temperature, humidity and vibration limits, thus they have the flexibility to remain powered for access and search of larger data spaces, as well as the ability to transmit longer data packets for simplified data retrieval. Also, they can power an internal real-time clock and apply an accurate time/date stamp to each recorded mobile sensor value or event.

The detailed data collected from the tags during intermodal container loadings and transportation may uncover inefficiencies in established procedures and among operations strategy elements that could not previously be identified, thus making its transportation processes more agile and safer and improve the overall quality of the general intermodal container transportation management information system, therefore, the efficiency of all transportation operations. Also, automatic tracking of information is valuable in many service operations: for many applications, it is sufficient to know that a tag has passed by a reader in a given location. The automatic wireless reading of multiple RFID tags creates an enormous data flow that is beneficial to the transport operation management of many transportation services, enabling improvements in the accuracy of delivery promise, and in the speed of cargo delivery, but hardens the part of that data analysis. Whereas, in an alert situation the source of the problem can be defined by some basic predefined business process rules within the basic transportation management information system, such that if an object passes into or out of a predefined secure area, or if a problem occurs during a cargo check, then this action can trigger also other events, processes, e-mail or SMS alerts or report notifications to occur automatically. Such safe precaution system would be capable of minimizing the time spent on cargo checks and would let the system automatically decide when to bother employees, thus minimizing the rate of errors in the proposed basic information system in real time manner.

This provides company managers with an up-to-the-minute picture of transportation processes and activities and that, in turn, allows them to respond to developing problem situations in a timely manner. Active RFID tags are also constantly powered, whether in range of a reader or not, and are therefore able to continuously monitor and record mobile sensor status, particularly valuable in measuring temperature, humidity and vibration limits, thus they have the flexibility to remain powered for access and search of larger data spaces, as well as the ability to transmit longer data packets for simplified data retrieval. Also, they can power an internal real-time clock and apply an accurate time/date stamp to each recorded mobile sensor value or event.

The detailed data collected from the tags during intermodal container loadings and transportation may uncover inefficiencies in established procedures and among operations strategy elements that could not previously be identified, thus making its transportation processes more agile and safer and improve the overall quality of the general intermodal container transportation management information system, therefore, the efficiency of all transportation operations. Also, automatic tracking of information is valuable in many service operations: for many applications, it is sufficient to know that a tag has passed by a reader in a given location. The automatic wireless reading of multiple RFID tags creates an enormous data flow that is beneficial to the transport operation management of many transportation services, enabling improvements in the accuracy of delivery promise, and in the speed of cargo delivery, but hardens the part of that data analysis. Whereas, in an alert situation the source of the problem can be defined by some basic predefined business process rules within the basic transportation management information system, such that if an object passes into or out of a predefined secure area, or if a problem occurs during a cargo check, then this action can trigger also other events, processes, e-mail or SMS alerts or report notifications to occur automatically. Such safe precaution system would be capable of minimizing the time spent on cargo checks and would let the system automatically decide when to bother employees, thus minimizing the rate of errors in the proposed basic information system in real time manner.

Conclusions

An approach for developing the interaction architecture of mobile devices and remote server systems with additional functionalities for contextual information transmission is proposed. The choosing of Petri nets allows describing the transportation system by imitational model and analyzing dynamic properties of this complex system. Petri nets provide effective formal means for description of decision making processes and scenarios of SIP communication protocol. For the semantic representation of data we are used class diagrams based of semantic object oriented model of UML. For risk representation and possibilities to evaluate the levels of risk we describe the set $S = \{ s_k \}$ of types of scenarios of accident events of transportation which we can to recognize. Scenarios are described using the probability of evolving of such type of scenarios. The proposed context modeling mechanism assures an always up-to-date context model that contains information on the transport device and location. We offer mobile internet services to extend the user interaction with architecture. The main advantage is the extensible architecture so that you can get the data to a mobile devises through web services. In this way, we try to solve the data integration of heterogeneous systems and compatibility issues.

References

[1] A. Shariat-Mohaymany and M. Babaei, An approximate reliability evaluation method for improving transportation network performance, *Transport* **25**(2), (2010), 193–202.

[2] J. C. Thill and H. Lim, Intermodal containerized shipping in foreign trade and regional accessibility advantages, *Journal of Transport Geography* **18** (2010), 530–547.

[3] E. D. Kreutzberger, Distance and time in intermodal goods transport networks in Europe: a generic approach, *Transportation Research Part* **A 42** (2008), 973–993.

[4] L. Ruiz-Garcia, P. Barreiro, J. Rodriguez-Bermejo, and J. I. Robla, Review. Monitoring the intermodal, refrigerated transport of fruit using sensor networks, *Spanish Journal of Agricultural Research* **5**(2), (2007), 142–156.

[5] W. Wen, An intelligent traffic management expert system with RFID technology, *Expert Systems with Applications* **37**(4) (2010), 3024–3035.

[6] C. Amador and J. P. Emond, Evaluation of sensor readability and thermal relevance for RFID temperature tracking, *Computers and Electronics in Agriculture* **73**(1) (2010), 84–90.

[7] M. Chavali, T. H. Lin, R. J. Wu, H. N. Luk, and S. L. Hung, Active 433 MHz-W UHF RF-powered chip integrated with a nanocomposite m-MWCNT/polypyrrole sensor for wireless monitoring of volatile anesthetic agent sevoflurane, *Sensors and Actuators* **A 141** (2008), 109–119.

[8] C. K. M. Lee and T. M. Chan, Development of RFID-based Reverse Logistics System, *Expert Systems with Applications* **36**(5) (2009), 9299–9307.

[9] R. Jedermann, C. Behrens, D. Westphal, and W. Lang, Applying autonomous sensor systems in logistics – combining sensor networks, RFIDs and software agents, *Sensors and Actuators* **A 132** (2006), 370–375.

[10] V. Pavliska, *Petri Nets as Fuzzy Modelling Tool*. (2006) Available from Internet: (http://irafm.osu.cz/research_report/84_rep84.pdf).

[11] D. Dzemydienė and R. Dzindzalieta, Development of decision support system for risk evaluation of transportation of dangerous goods using mobile technologies. In: M. Grasserbauer, L. Sakalauskas, E. K. Zavadskas, editors, *Knowledge-based Technologies and OR Methodologies for Strategic Decisions of Sustainable Development*, 2008, 108-113.

[12] A.A. Bielskis, D. Dzemydienė, V. Denisovas, A. Andziulis, and D. Drungilas, An approach of multi-agent control of bio-robots using intelligent recognition diagnosis of persons with moving disabilities, *Technological and Economic Development of Economy* **15**(3)(2009), 377-394.

[13] N. Batarlienė and A. Baublys, Mobile solutions in road transport, *Transport* **22**(1) (2007), 55-60.

[14] N. Batarlienė, Implementation of advanced technologies and other means in dangerous freight transportation, *Transport* **22**(4) (2007), 290–295.

[15] B. Fabiano, F. Currò, A. P. Reverberi, and R. Pastorino, Dangerous good transportation by road: from risk analysis to emergency planning, *Journal of Loss Prevention in the Process Industries* **18**(4-6) (2005), 403-413.

[16] B. Fabiano, E. Palazzi, F. Currò, and R. Pastorino, Risk assessment and decision-making strategies in dangerous good transport. In: *From an Italian case-study to a general framework. Loss Prevention and Safety Promotion in the Process Industries,* Proceedings 2. Elsevier Science, Amsterdam, 2001, 955-966.

[17] A. Prekopa, *Stochastic Programming*. Kluwer, 1995.

[18] P. Leonelli, S. Bonvicini, and G. Spadoni, Hazardous materials transportation: a risk-analysis-based routing methodology. *Journal of Hazardous Materials* **71**(1) (2000), 283-300.

[19] G. F. List, P. B. Mirchandani, M. Turnquist, and K. G. Zografos, Modeling and analysis for hazardous materials transportation. Risk analysis, routing/scheduling and facility location. *Transportation Science* **25**(2) (1991), 100-114.

[20] D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, Ch. Ferris, and D. Orchard, *Web Service Architecture.* W3C Working Group Note 11 February 2004. W3C, 2004. Available from: http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/.

[21] D. Dzemydienė, S. Maskeliūnas, and K. Jacobsen, Sustainable management of water resources based on web services and distributed data warehouses, *Technological and Economic Development of Economy* **14**(1) (2008), 38-50.

[22] *eSafety Initiative. Safer Roads for European Citizens*, 2005-2007. Available from: http://www.esafetysupport.org/en/esafety_activities/.

[23] M. Gudgin, M. Hadley, and T. Rogers, *Web Services Addressing 1.0 – Core*. W3C Recommendation 9 May 2006. W3C, 2006. Available from: http://www.w3.org/TR/ws-addr-core/.

[24] *Hazardous Substances Ordinance – GefStoffV of 23 December 2004*. Available from: www.zuv.uni-eidelberg.de/sw/gefahrstoffe/gesetze/HazardousSubstances-Ordinance.pdf.

[25] K. Šutienė, D. Makackas, and H. Pranevičius, Multistage k-means clustering for scenario tree construction, *Informatica* **21**(1) (2010), 123–138.

[26] C. M. Huang and C. H. Lee, Signal reduction and local route optimization of SIP-based network mobility. In: *Proceedings of the 11th IEEE Symposium on Computers and Communications* (ISCC), 2006, 482–87.

[27] C. M. Huang, C. H. Lee, and J. R. Zheng, A novel SIP-based route optimization for network mobility. IEEE *Journal on Selected Areas in Communications* **24**(9) (2006), 82–91.

[28] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collons, *GPS – Global positioning system. Theory and practice.* Springer, Wien, 1997.

[29] D. Johnson, C. Perkins, and J. Arkko, *Mobility Support in IPv6*. IETF RFC 3775, 2004.

[30] S. Kreiensen and K. Kyamakya, *Project mobiGPS*, 2004.

[31] H. Mahmoud, *J2ME and Location-Based Services*, 2004.

[32] *SIP: Session Initiation protocol*, RFC 3261. Network Working Group, The Internet Society, June 2002. Available from: http://www.ietf.org/rfc/rfc3261.txt.

[33] J. White and D. Hemphill, *Java 2 Micro Edition. Java in Small Things*. Manning Publications Co., 2002. Available from: http://www.desifile.com/files/27615241809/Manning - Java 2 Micro Edition.pdf.

Formulating the Enterprise Architecture Compliance Problem

Vytautas ČYRAS[a] and Reinhard RIEDL[b]
[a] *Vilnius University, Vilnius, Lithuania*
[b] *Bern University of Applied Sciences, Bern, Switzerland*

Abstract. We start from the Klaus Julisch's (2008) IT compliance problem definition and make an attempt to formulate the enterprise architecture compliance problem (EACP). The challenging issues comprise the complexity of the law phenomenon, compliance frameworks and methodologies to check EA for non-compliance with laws and regulations. We hold that a compliance methodology should take into account "shared" relevant laws and a requirements engineering framework. We reflect mainly on the view of enterprise architects on legal informatics and a vision driven approach on requirements elicitation in the context of enterprise engineering, which was proposed by Albertas Čaplinskas (2009). Then we raise a question of placing EACP into the Bonazzi-Hussami-Pigneur regulation and IT alignment framework (2009).

Keywords. Regulatory compliance, electronic identity compliance, enterprise systems, transparency engineering, legal informatics, legal visualization

Introduction

This paper attempts to overview some models which could contribute to formulate regulatory compliance problems. However, the field appears too broad to master with a sweep of the arm. A unified "enterprise-wide" compliance process remains an ambition. Thus the authors present reflections on various issues. The message is that a compliance methodology should follow a requirements engineering framework because the latter combines business, IT and legal perspectives.

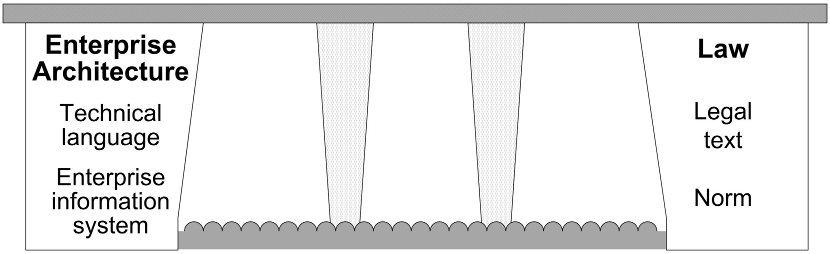


Figure 1. Building a bridge between enterprise architecture and law

This paper is challenged by a (naïve?) question: Does an enterprise architecture comply with the law? It can be compared with Alan Turing's "Can machines think?" [21]. Both cannot be answered simply yes or no. Each challenges to formulate a distinguished problem. Turing replaces the question with the following series: What does it mean 'think', 'machine' and 'can'? This was discussed in the early artificial intelligence. Similarly we can ask: What enterprise architecture is? and Which law?

The "naive" compliance problem formulation above is similar to bridging enterprise architecture and law (Figure 1). The bridge metaphor is generally used in knowledge visualization. We utilize it because we see a similarity with the bridge between informatics and law, which was proposed by Friedrich Lachmayer to characterize the subject matter of legal informatics [5]. One arch is hardly feasible and therefore the bridge consists of multiple arches. Thus a multiphase transformation process emerges. The transformation is about bridging legal requirements and the enterprise information system.

Transparency optimization is a major purpose in EA. Legal requirements are one of a kind among all requirements tackled by enterprise architects. Different legal issues are concerned in every EA perspective. The enterprise architect's perspective has the task to integrate all the different views on EA, in particular, the business view, the ICT view, and the legal view. Figure 2 shows the key terms within the point of departure.

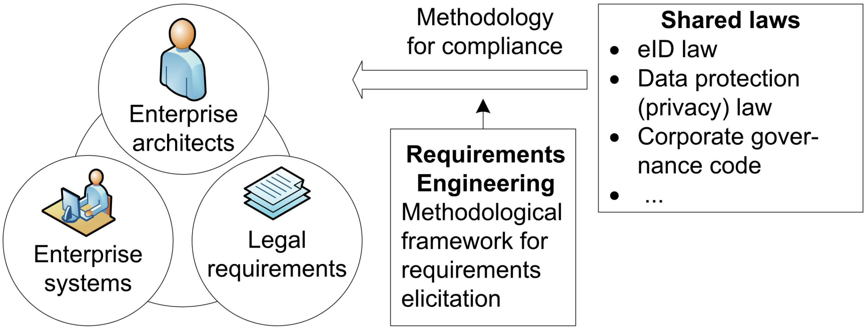

Enterprise architects
Enterprise systems
Legal requirements
Methodology for compliance
Requirements Engineering
Methodological framework for requirements elicitation
Shared laws
• eID law
• Data protection (privacy) law
• Corporate gover-nance code
• ...

Figure 2. The key concepts tackled in this paper

1. Motivating the Research

1.1. e-Identity and e-Banking Requirements within the STORK 2.0 Project

A practical motivation for academia can be illustrated by the STORK 2.0 project[1] that concerns the design and implementation of the foundations for a unified European identification and authentication space. STORK (Secure idenTity acrOss boRders inKed) [2008-2011] established a European electronic identity (eID) interoperability platform which allows holders of national eIDs to access cross-border electronic services within six pilot applications. Its extension STORK 2.0 [2012-2014] will extend the range of services within four new pilots. One of them is focused on e-banking, mainly retail banking, the scene for secure e-invoicing. The use cases specify three distinct cross-border scenarios using national eIDs: (1) opening a bank account online using eIDs for identity authentication, (2) logging into a cross-border e-banking portal using eIDs, and (3) authorizing payment of electronic bills using eIDs.

[1] Cofunded under EU ICT Policy Support Programme as part of the Competitiveness and Innovation framework Programme (CIP); see also https://www.eid-stork.eu/.

The big picture for the whole project aims at a shift from interoperability for a few national eIDs to the creation of a single European identity space for borderless e-business. All trustworthy eIDs, both national and commercial, should be usable in this space upon the accreditation of the eID providers according to a 4-level quality standard for the trustworthiness of eIDs.

Non-repudiation is a critical requirement for electronic business. It is also a precondition for compliance as otherwise transactions could be repudiated. Both can be based on the authentication of interaction partners and a trustworthy recording of interaction activities. Depending on the criticality of business issues, different levels of quality should be met by the authentication procedures corresponding to different levels of quality for the eIDs used for authentication. But the vision is that it is only the quality level which decides on the width of use of an eID.

The feasibility and the value of the design for a single European identification and authentication space will be validated within STORK 2.0 through four pilots. One pilot is concerned with e-banking. Its key vision is to move identity and access management (IAM) out of the core banking IT system. Authentication should be possible with any eIDs issued by accredited eID providers guaranteeing the highest quality level of trustworthiness. Of course, a solution must include major national electronic identities in Europe, if they comply with highest quality standards. However, even for them legal issues are unclear.

Following is a use case to check for compliance. A company representative with an eID from country *X* (e. g. Germany), working in a company from country *Y* (e.g. Switzerland) logs into a banking platform in country *Z* (e.g. Austria, Lithuania or US). The number of potential customers comprises foreign nationals, for instance, those living in Switzerland and cross border commuters.

Do the requirements for e-banking comply with national eID laws? Can the proof of identity be transferred from the issuing of an eID be transferred to the opening of a bank account with this eID? In many countries this is an open question. For sure, some conflicts exist [14, p. 440]:

> Seen from a European political perspective, eIDs are primarily in potential conflict with privacy protection rights and thus with data protection laws. However, seen from a broader political perspective, the design of a single identification and authentication space also touches the so far hardly discussed eventual right for being recognized by electronic services.

This indicates that compliance is a tricky and much broader issue than it appears to be at first glance. For example, excluding some users may be legally compliant in one country and clearly non-compliant in another.

The STORK 2.0 project also tackles common infrastructure for federated e-Government, in particular in Switzerland. Today's challenges are:

- organizational and business models,
- implementation of a government cloud,
- refinement of the existing enterprise architecture in order to get it "working".

Tomorrow's challenge is enterprise architecture design for the implementation of the Lenk-Schaffroth-Schuppan vision of networked government, which links processes across different public administration organizations [12]. Future challenge is the separation of distribution, execution, and control in order to implement shared service centres for core tasks of the state. Apart from other challenges, there highly complex compliance issues to be considered.

As a final (much more simple) example, we may compare implementations for one-stop government, which depend on the integration of truly independent processes in different government agencies. For such implementations the choice of a tier for integration is critical. It makes a big difference whether it is done in the web-tier (front office integration) or in the application tier (back-office integration, e.g. with WS-BPEL, Web Services Business Process Execution Language). The latter is in many cases not legally compliant because it does not guarantee the immediate registration of incoming requests at every organization.

1.2. Formulating the Enterprise System Compliance Problem

Klaus Julisch (2008) suggests academia a paradigm shift: from "selling" security when organizations seek to "buy" compliance to complementing current security research by additional research into security compliance:

> [A]s long as careers are terminated and people go to jail…for failures in compliance – rather than security – the commercial world will continue to pursue compliance rather than security as their primary goal. [8, p. 71]

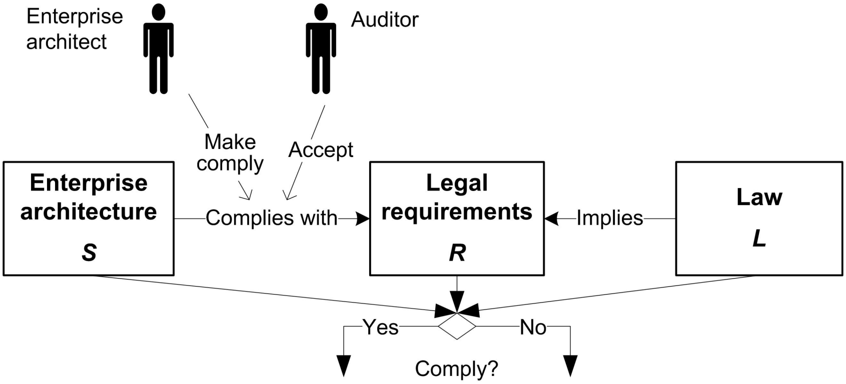


Figure 3. Enterprise system compliance problem

Julisch defines: "security compliance, in IT systems, is the state of conformance with externally imposed functional security requirements and of providing evidence (assurance) thereof." [8, p. 72] He defines the security compliance problem as follows:

> *Definition*: Given an existing IT systems *S* and an externally imposed set *R* of security requirements. The *Security Compliance Problem* is to make system *S* comply with the security requirements *R* and to provide assurance that an independent auditor will accept as evidence of the compliance of system *S* with requirement *R*.

Following the definition above, we would formulate the *Enterprise Architecture Compliance Problem* (EACP). It is (1) to make enterprise architecture *S* comply with requirements *R* that relate to a law *L*, and (2) to provide assurance that an independent auditor will accept this as evidence (Figure 3).

We simply added a law *L* to Julisch's formulation. Semiformal definitions above can only serve as a first iteration. Problem solutions in practice can hardly result with yes or no. Practice involves more elements. Feedback loops would improve *S*, *R* and *L*. Conceptualisations of *L* may involve different elements depending on abstraction level. A legal principle, a whole statute or a specific provision may stand for *L*.

1.3. Complexity Issues when Attempting to Formalise the Law in the Context of EA

Failure to understand the law is one of non-compliance reasons [18, p. 59–61]. This failure can be examined from the management perspective. We comment further from the legal perspective. The texts of laws constitute only a part of the whole legal system. The meaning (*Sinn*) of law – the Ought realm – can hardly be understood from the sole legal text. Therefore a freshman can hardly understand the spirit of law while reading a separate statute. On the other hand, the compliance problem can hardly be reduced to tick the box. The law does not allow to be easily represented for EA developers whose purpose is to enforce the law. The following issues raise difficulties:

1. *Abstractness of norms.* Norms are formulated (on purpose) in abstract terms.
2. *Principle vs. rule.* The difference in regulatory philosophy between the US and other countries [18, p. 46].
3. *Open texture.* This can be illustrated by H. L. A. Hart's example of "Vehicles are forbidden in the park".
4. *The myriad of regulatory requirements.* Compliance frameworks are multidimensional.
5. *Heuristics.* High level concepts are translated into invented low level ones.
6. *Teleology.* The purpose of a legal norm usually can be achieved by a variety of ways. They need not to be listed in a statute and specified in detail.
7. *Legal interpretation methods.* The meaning of a legal text cannot be extracted from the sole text. Apart from the grammatical interpretation, other methods can be invoked, such as systemic and teleological interpretation.
8. *Consciousness of the society.* Modeling it is a tough task.

2. A Variety of Factors to Comply with

Note that judges are allowed to have different opinions. Are auditors allowed too? The COSO framework[2] was issued prior to Sarbanes-Oxley Act of 2002 (SOX)[3]. Deterring fraudulent financial reporting is an aim. The use of the COSO framework by company management shows the scale from 'no extent' to 'large extent' [11]. The Directive 2008/30/EC[4] can be compared with the impact of SOX in the US. Information technology internal controls are not an exclusive concern of COSO.

Anthony Tarantino (2008) devotes the whole book to Governance, Risk and Compliance (GRC). He suggests taking a holistic approach. In particular, he addresses the risk concept [20, p. 15–17, 236–237]. Banking's categorization accords[5] describe seven major areas of operational risk:

1. Internal fraud: unauthorized activities; theft and fraud.
2. External fraud: external security; theft and fraud.

[2] The Committee of Sponsoring Organizations (COSO) framework was originally issued in 1992 (entitled Internal Control – Integrated Framework) and updated in 2004 (Enterprise Risk Management), see http://www.coso.org/IC-IntegratedFramework-summary.htm.

[3] See Wikipedia, "Sarbanes–Oxley Act", http://en.wikipedia.org/wiki/Sarbanes%E2%80%93Oxley_Act.

[4] Directive 2008/30/EC of the European Parliament and of the Council of 11 March 2008 amending Directive 2006/43/EC on statutory audits of annual accounts and consolidated accounts. In Lithuania see the Audit Law of 2011-11-17 (originally 15-06-1999) VIII-1227, Gazette 1999, Nr. 59-1916.

[5] See Wikipedia, "Basel II", http://en.wikipedia.org/wiki/Basel_II.

3. Employment practices: employee relations; safe environment; diversity and discrimination.
4. Clients, products, and business processes: suitability, disclosure, and fiduciary aspects; product flaws; improper business or market practices; advisory activities; selection, sponsorship and exposure.
5. Damage to physical assets: disasters and other events.
6. Business disruptions and system failures: systems.
7. Execution, delivery, and process management: transaction capture, execution, and maintenance; monitoring and reporting; incomplete legal documentation; customer account management.

This Level 1 and Level 2 categorization can serve as a framework. It illustrates a variety of risk factors, which have to be faced by auditors and other personnel.

Compliance is a multi-criteria problem. A single framework or standard can hardly be a solution to all compliance and control needs:

> Absolute adherence to a regulation by adopting a basic framework without considering the entire organization and threats that affect it make the organization compliant, but not secure or resilient to operational disruptions. [6, p. 62]

3. Elements of Enterprise System

According to the systems engineering view, an enterprise system consists of three subsystems [4] listed below. However, there is no generally accepted agreement.

1. *Enterprise business system*. It is comprised of business actors, resources and business processes;
2. *Enterprise information system* (IS) "is a whole formed out of organisational memory and sets of information processing actors (IPA), information flows, and interrelated information processing processes implemented in accordance with the enterprise information processing policies and standards" [4]
3. *Enterprise application system*. It is comprised of hardware agents, protocols, knowledge bases and software application programs.

Other elements can be distinguished, too. This depends on an author's view. Ross et al. note that the term 'architecture' has acquired a negative connotation in some companies and quote saying "Architectures, like fondue sets…, are rarely used." [17, p. 47] They make emphasis on distinguishing between enterprise architecture and IT architecture. They also note that the IT unit typically addresses four levels of enterprise architecture [17, p. 48–49]:

1. *Business process architecture*. The activities or tasks composing major business processes identified by the business process owners.
2. *Data or information architecture*. Shared data definitions.
3. *Applications architecture*. Individual applications and their interfaces.
4. *Technology architecture*. Architecture services and the technology standards they are built on.

Subsystems and systems thinking are stressed in [7, p. 29–52]. First a *system* is defined as "a set of discernable, interacting parts or subsystems that form an integrated whole that acts with a single goal or purpose" (p. 29). Then EA is characterised:

> An *Enterprise Architecture* describes the structure of an enterprise, its decomposition into subsystems, the relationships between the subsystems, the relationships with the external environment, the terminology to use, and the guiding principles for the design and evolution of an enterprise. [7, p. 102]

One of the problems right now with enterprise architecture (see [17]) is that for reasons of simplicity those practically usable in real life focus on very few aspects of a real world IS, usually issues which are shared throughout the organization. Depending on the operating model of the organization, this is just technology, or in addition data and/or business processes. In the spirit of EA in the sense of Ross et al., it would make sense to define the "shared" relevant laws and integrate them; cf. [17, p. 12–13]:

> Companies are buffeted by constant changes in regulations, such as Sarbanes-Oxley, Basel II, and HIPAA. As companies become global, they become accountable for increasingly complex reporting requirements. …Companies may not be able to anticipate new regulations, but they can increase the likelihood that needed data is readily available or can easily be accumulated.

Ross et al. suggest encapsulating enterprise architecture in a core diagram, which depicts a high-level view of the process, data, and technologies constituting the desired foundation for execution. Here we raise a question: how to formulate the EA compliance problem once such a one-page core diagram is provided? Writing a list of compliance requirements? An answer should concern a concluding remark that enterprise architecture is not a detailed blueprint of systems, data, and technology, but instead a business vision [17, p. 206].

Enterprise architects check the architecture for potential conflicts with the law. The regulations which influence enterprise architectures, perhaps SOX, can be barely aware [13]. The following relationships can be identified here:

- Architecture descriptions have to leverage checking compliance.
- Legal informatics experts can contribute to legislation, esp. in e-Government.
- Enterprise architects become important partners for legal informatics experts. This is possible in the revision of the law, e.g. in digital identity regulations.
- Contacts with authorities when anticipating ICT perspectives. Regulation of software exchange, for instance, modules in finance informatics.
- Ideas in legal informatics; patterns and anti-patterns.

4. Enterprise Compliance Process

Financial compliance process is an important but not the sole issue of conformance. Enterprise content management (ECM) systems are focused in [9] and a high-level compliance process is provided; see Figure 4. Following is the list of standard requirements that any ECM vendor should provide: library services, repository search, document routing, central user administration, support for all popular text file formats, document imaging. More complex requirements: document-centric collaboration, compound documents support, digital assets management, records management, rule-driven workflow, process management, advanced security, etc. [9, p. 262–264].

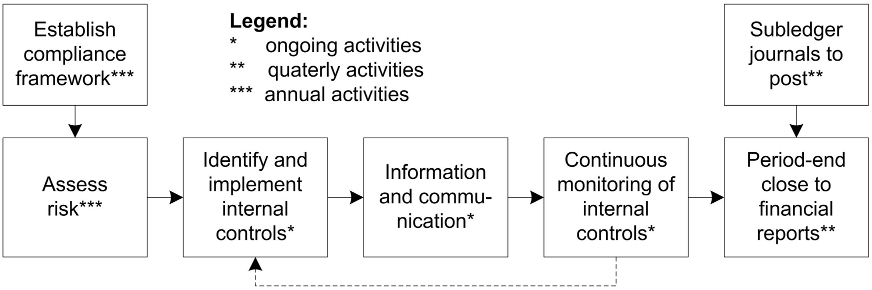

Establish compliance framework***
Legend:
* ongoing activities
** quaterly activities
*** annual activities
Subledger journals to post**
Assess risk***
Identify and implement internal controls*
Information and commu-nication*
Continuous monitoring of internal controls*
Period-end close to financial reports**

Figure 4. High-level compliance process [9, p. 260]

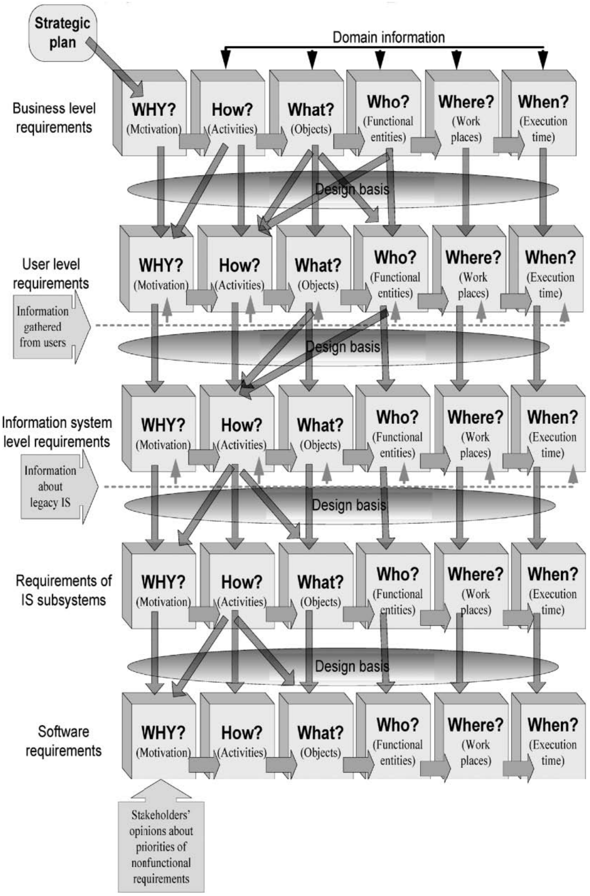

Strategic plan
Domain information
Business level requirements
User level requirements
Information gathered from users
Information system level requirements
Information about legacy IS
Requirements of IS subsystems
Software requirements
WHY? (Motivation)
How? (Activities)
What? (Objects)
Who? (Functional entities)
Where? (Work places)
When? (Execution time)
Design basis
Stakeholders' opinions about priorities of nonfunctional requirements

Figure 5. Čaplinskas' methodological framework for requirements elicitation, analysis, specification and validation; see [4]

5. The Legal Perspective in Enterprise Engineering

The only true purpose of the work of enterprise architects is transparency optimization in an organization. Three central perspectives to enterprise systems can be concerned:

1. business perspective,
2. ICT (information and communication technologies) perspective,
3. legal perspective.

Continuing the list above, other potential perspectives can be mentioned:

4. internal communication perspective (a direction tool),
5. public relations and marketing perspective,
6. political and economic perspective (probably in e-Government).

In our research we focus on the legal perspective. More perspectives are concerned in architectural frameworks such as Zachman's one [19]. Zachman's idea to decompose the system into a number of perspectives and focus areas has served as a theoretical basis for the vision driven approach proposed by Čaplinskas. Zachman decomposes each perspective into six focus areas to be answered: what (data)? how (function)? where (network)? who (people)? when (time)? and why (motive)? The following five perspectives (views, levels) are shown in Figure 5:

1. business level requirements (the view of business analyst),
2. user level requirements (the view of stakeholders),
3. IS (information system) requirements (the view of IS analyst),
4. the requirements of IS subsystems (the view of IS engineer),
5. software requirements (the view of software analyst).

Other perspectives, which are out of the Zachman framework (that is an architectural one), are commented in [4, p. 355]:

> To be complete, it should additionally include the requirements of software components (the view of software architect), the implementation requirements (the view of software engineer), the process requirements (the view of process engineer), and the testing requirements (the view of tester). …The first five perspectives differ from corresponding ones provided by the Zachman's framework because they are designed for different purposes.

Each perspective (level) presents a model of the system. Each phase of system's life-cycle is subject to technical standards. The concepts of a to-be-system and requirements are related to law. The requirements document (system specification) is part of the contract with a customer. Every requirement is based on a norm. This norm is present in a technical standard, business rule or other kind of legal source. The difference in the nature of requirements stems from the difference of norms.

6. Towards a Methodology of the Compliance with the Law

The end-to-end enterprise architecture compliance problem is too large and too complex for any one company to tackle. Similar is with compliance auditing, frameworks and good practices. Following are theme aggregates to shape the integration of different recourses compliant with the law [13]:

1. internal arrangement of transparency,
2. methods for the legal architecture view as part of enterprise architecture,
3. design methods for law-triggered changes in the enterprise architecture.

We think that Čaplinskas' approach could provide a framework to shape the methods above. First, it is vision driven. Second, a legal perspective can be added. Further steps face the following problems. An analyst can hardly be aware of legal norms in different branches of law. Therefore methodologies are needed. A trivial idea might be simply to check the requirements in each cell in Figure 5 for compliance. This would be classified as an ex-ante solution [2] "to design an artefact aimed at avoiding actions that are not compliant". However to lower the risk of violating strategic alignment, a holistic approach has to be undertaken. Ex-post solutions are "to design an artefact to assess the level of compliance".

Risk management for information technology is a growing challenge for GRC [20, p. 18]. The National Institute of Standards and Technology (NIST) provides IT guidelines for U.S. federal agencies. NIST's Special Publication 800-30 (Risk Management Guide for Information Technology Systems) recommendations involve three basic processes: risk assessment, risk mitigation, and evaluation and assessment. Absolute control measures are often cost prohibitive and require IT professionals to weigh the cost versus benefits. This process is complicated by the hundreds of software tool suppliers promising to fix their GRC problems [20, p. 19].

The Bonazzi-Hussami-Pigneur IT compliance framework is worth a special attention. Two dimensions, *Legal* and *IT*, and two kinds of sources of regulation to comply with, *External* and *Internal*, are depicted with squares (Figure 6). Distinguished alignments are represented with arrows. A direction points to the defined artefact [2].

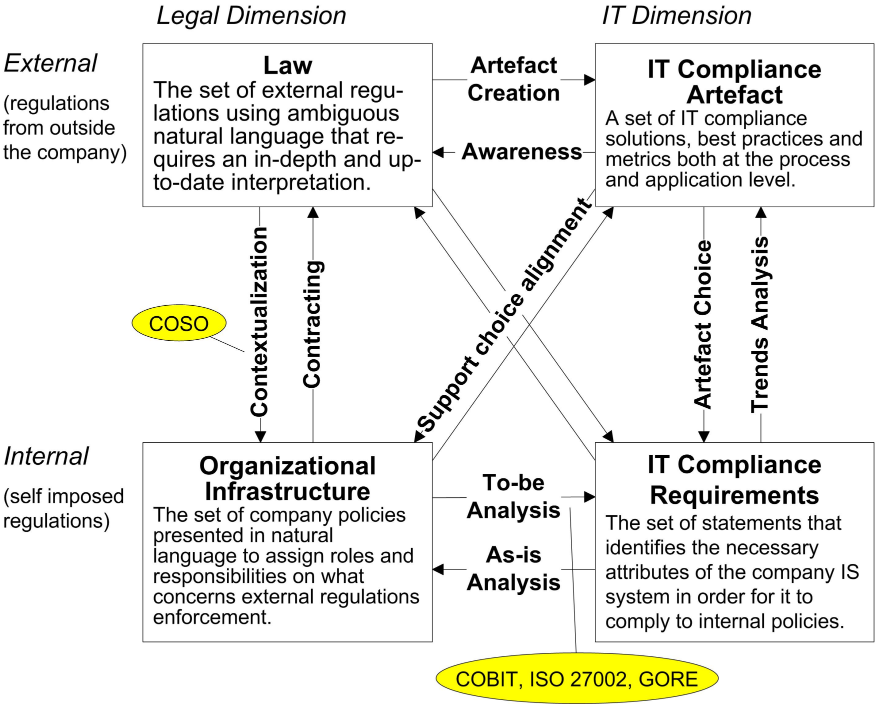

Legal Dimension
IT Dimension
External
(regulations from outside the company)
Law
The set of external regulations using ambiguous natural language that requires an in-depth and up-to-date interpretation.
Artefact Creation
Awareness
IT Compliance Artefact
A set of IT compliance solutions, best practices and metrics both at the process and application level.
Contextualization
Contracting
COSO
Support choice alignment
Artefact Choice
Trends Analysis
Internal
(self imposed regulations)
Organizational Infrastructure
The set of company policies presented in natural language to assign roles and responsibilities on what concerns external regulations enforcement.
To-be Analysis
As-is Analysis
IT Compliance Requirements
The set of statements that identifies the necessary attributes of the company IS system in order for it to comply to internal policies.
COBIT, ISO 27002, GORE

Figure 6. The Bonazzi-Hussami-Pigneur regulation and IT alignment framework; adapted from [2]

Every concept in Figure 6 denotes a broad field. Corporate noncompliance, corruption, etc. are just a few examples of violations. Noncompliance can be civil, criminal, administrative, but also reputational or market based. To analyse for conformance, one analyst can hardly be aware of norms in various branches of law.

To-be Analysis can be treated in different ways, depending on controls or IT risks

(e. g. COBIT[6], ISO 27002[7], GORE [15]). COBIT concerns IT governance and best practices. COBIT view of the implementation of a system's infrastructure can be summarized as follows: plan and organize, acquire and implement, deliver and support, and monitor and evaluate [16, p. 322]. Like its COSO control counterpart, it is the framework for the management of IT processes.

The design of law-compliant information systems is an area on a lower level of abstraction, though falls within the framework above. As noted in [10], representation of legal requirements were lacking in early works. Therefore they propose a framework to model the area which is structured in three dimensions: (1) field of law (flow control, reporting, web applications, etc.), (2) modelling level (analysis level, model level, metamodel level), and (3) research goal (explanation- or design-oriented research).

An interesting perspective for the actual implementation of such a holistic approach is provided by agile software development [1]. Agility in IS solution development was first introduced in the extreme programming (XP) movement. It is nowadays extended to many IS related tasks and applied even in development and design tasks without much IS relationship but with a general multi-disciplinary and multi-stakeholder challenge. The key two ideas of agility is to involve all expertises needed for a good IS solution in the team working on the solution and to develop the solution in short cycles, where at the beginning of each cycle a requirements analysis takes place and at the end of each cycle a working (running) in-between-product is delivered to users. This original concept has been adapted to more conceptual contexts like enterprise architecture management, where no running products can be delivered. In the original setting the key challenge is to design a good basic overall structure, as the spirit of agility contradicts the planning. However, in principle agility proposes an interesting split for a holistic approach. The integration of multidisciplinary perspectives is split into a very rudimentary strategic core concept and a concrete teamwork, where requirements analysis and holistic solution development is done iteratively.

7. Conclusions

The paper presents reflections on different issues of compliance. The authors are influenced by both formal models which are used in computer science and descriptive methods of social sciences (including law). This is all for the best in the present research. However, there is no silver bullet to attack regulatory compliance requirements – no one-off, best-of-breed solution. Similar is with the theoretically formulated Enterprise Architecture Compliance Problem (EACP). Positioning it in an IT alignment framework is a challenge. Though various compliance processes are positioned differently even within the two dimensions of IT and law. Requirements engineering contributes combining business, IT and law.

EACP formulation deals with static artefacts. The idea is to restrict with ex-ante analysis. Dynamics such as management loops and internal controls, which are usually

[6] COBIT is a framework for IT governance and control. See Wikipedia http://en.wikipedia.org/wiki/COBIT and http://www.isaca.org/Knowledge-Center/cobit/Pages/Overview.aspx.

[7] An information security standard, entitled *Information technology – Security techniques – Code of practice for information security management*; see Wikipedia, ISO/IEC 27002 http://en.wikipedia.org/wiki/ISO_17799. ISO 27001 is entitled *Information technology – Security techniques – Information security management systems – Requirements*. It is a specification. It uses words like 'shall'. ISO 27002 is a code of practice, not a specification, cf. [3].

involved in ex-post approaches, would make the formulation too complex to master. Process models involve costs-benefits tradeoffs, for example, fast, lower quality, general but cheap process versus slow, higher quality, specific but expensive one.

Acknowledgements

We are grateful to Friedrich Lachmayer for the encouragement of this joint research. Thanks also to anonymous reviewers for shaping the paper.

References

[1] K. Beck et al., *Manifesto for Agile Software Development*, Agile Alliance, 2001. Available from: http://agilemanifesto.org/ retrieved 11-06-2012.

[2] R. Bonazzi, L. Hussami, and Y. Pigneur, Compliance management is becoming a major issue in IS design. In: A. D'atri, D. Saccà, editors, *Information Systems: People, Organizations, Institutions, and Technologies*, Springer, 2009, 391–398.

[3] A. Calder, ISO 27001 and ISO 17999. In: A. Tarantino, editor, *Governance, Risk, and Compliance Handbook*, John Wiley, 2008, 169–179.

[4] A. Čaplinskas, Requirements elicitation in the context of enterprise engineering: a vision driven approach, *Informatica* **20**(3) (2009), Lithuanian Academy of Sciences, 343–368.

[5] V. Čyras and F. Lachmayer. Multiphase transformation in the legal text-to-program approach. Submitted to: I. Takashi, H. Mori, editors, *Liber amicorum Guido Tsuno*. Chuo University, Japan, 2012.

[6] J. J. Deluccia, *IT Compliance and Controls: Best practices for Implementation*, John Wiley, 2008.

[7] R. E. Giachetti, *Design of Enterprise Systems: Theory, Architecture, and Methods*, CRC Press, 2010.

[8] K. Julisch, Security compliance: the next frontier in security research. In: *Proceedings of the 2008 workshop on New security paradigms NSPW'08*, ACM, 2008, 71–74.

[9] J. Koo, What to look for in enterprise content management for compliance. In: A. Tarantino, editor, *Governance, Risk, and Compliance Handbook*, John Wiley, 2008, 259–266.

[10] R. Knackstedt, M. Heddier, J. Becker, Fachkonzeption Rechtskonformer Informationssysteme als Anwendungsgebiet der Rechtsvisualisierung. In: E. Schweighofer, F. Kummer, editors, *Europäische Projektkultur als Beitrag zur Rationalisierung des Rechts, Tagungsband IRIS 2011*, OCG, 549–558.

[11] T. Leech, COSO – is it fit for purpose? In: A. Tarantino, editor, *Governance, Risk, and Compliance Handbook*, John Wiley, 2008, 65–75.

[12] K. Lenk, T. Schuppan, and M. Schaffroth, *Networked Public Administration: Organisational Concept for a Federal eGovernment Switzerland*. eCH-0126 White Paper. Federal Department of Finance FDF, Switzerland, 2010. Available from: http://www.ech.ch.

[13] R. Riedl, Rechtsinformatik aus Sicht des Unternehmensarchitekten. In: A. Geist, C. R. Brunschwig, F. Lachmayer, G. Schefbeck, hrsg. *Strukturierung der Juristischen Semantik – Structuring Legal Semantics. Festschrift für Erich Schweighofer*, Editions Weblaw, Bern, 2011,257–269.

[14] R. Riedl et al., A multiperspective view of unified identification and authentication spaces. In: E. Schweighofer, F. Kummer, W. Hötzendorfer, editors, *Transformation of Legal Languages, Proceedings of IRIS 2012*, OCG, Wien, 2012, 439–443. Available from: http://jusletter-eu.weblaw.ch/.

[15] A. Rifaut and E. Dubois, Using goal-oriented requirements engineering for improving the quality of ISO/IEC 15504 based compliance assessment frameworks. In: *16th IEEE International Requirements Engineering Conference, RE 2008*, 33–42, IEEE Computer Society, 2008.

[16] I. Rodgers, Internal controls best practices. In: A. Tarantino, editor, *Governance, Risk, and Compliance Handbook*, John Wiley, 2008, 301–323.

[17] J.W. Ross, P. Weill, and D. C. Robertson, *Enterprise Architecture as Strategy: Creating a Foundation for Business Execution*, Harward Business School Press, Boston, 2006.

[18] M. Silverman, *Compliance Management for Public, Private, or Nonprofit Organizations*, McGraw-Hill, New York, 2008.

[19] J. Sowa and J. Zachman, Extending and formalizing the framework for information systems architecture, *IBM Systems Journal* **31**(3) (1992), 590–616.

[20] A. Tarantino, editor, *Governance, Risk, and Compliance Handbook: Technology, Finance, Environmental, and International Guidance and Best Practices*. Introduction, 1–37. Operational risk management in financial services, 233–256. John Wiley, 2008.

[21] A. Turing, Computing machinery and intelligence, *Mind* **59** (October 1950), 433–460.

Investigation of Data Transfer Capabilities for Heterogeneous Service Support in Critical Mobile Objects Communication Situations

Mindaugas KURMIS[a,c,1], Dale DZEMYDIENE[b] and Arunas ANDZIULIS[c]
[a] *Vilnius University, Institute of Mathematics and Informatics, Lithuania*
[b] *Mykolas Romeris University, Department of Informatics and Software Systems Lithuania*
[c] *Klaipeda University, Informatics Engineering Department, Lithuania*

Abstract. Research of heterogeneous service providing in the fast-changing topology vehicular communication networks are important because expansion and integration of this intelligent transport systems platform would greatly improve traffic safety and reduce injuries on the road. At the same time, the trips would be more comfortable. In this work, it is investigated data-transfer capabilities for heterogeneous service support, road safety, assessed their integration potentials and prospects in vehicle communication networks with changing topology. It is showed that to provide quality heterogeneous services it is necessary new routing protocols and channel access methods for the large volume fast changing topology networks.

Keywords. Multimedia services, vehicular communication networks, routing ad-hoc networks, mobile nodes, changing topology

Introduction

Today, the vehicle is a very important component of human life, so installed intelligence based software and hardware equipment, can improve the level of travel safety and comfort. Currently, one of the most attentions attracting mobile communication technology is vehicular wireless communication networks. They offer the potential to develop and produce safer, more reliable, economic and comfortable vehicles. These networks are gaining more and more commercial relevance, since the adoption of DSRC (Dedicated Short-Range Communication) / IEEE 802.11p (Wireless access in vehicular environments (WAVE)) standards in both the EU and the U.S., given the possibility to reach an entirely new level of service in a vehicle, covering many areas, including road safety, traffic management, comfort applications. Vehicles do not have strict restrictions on power consumption, and therefore, can be easily equipped with powerful computing devices, wireless transmitters, sensors, complex

[1] Mindaugas Kurmis, Vilnius University, Institute of Mathematics and Informatics, Akademijos str. 4, Vilnius, Lithuania, LT-08663, mindaugas.kurmis@mii.vu.lt, +37065370031

systems - GPS, photo / video cameras, vibration, acoustic, chemical sensors and, etc. [1].

Practices of vehicular communication network's deployment, research and scientific projects are developing in two directions: direct vehicle-vehicle (V2V) communication and vehicle-to-infrastructure (V2I) communication [2]. Research in this area addresses many complex communication problems as there are many specific determinants of the quality of communication, including highly dynamic traffic and communication conditions, frequent disconnection of nodes as well as heterogeneity of data transmission links.

This paper explores the evaluation of the data-transfer efficiency in a mobile communication network when the sender and the receiver is moving in opposite directions at high speed. It is organized as follows: in Section 1, we analyze the vehicular communication networks and their architecture, in Section 2, we briefly present related works. Section 3 describes the experiment methodology and simulation model. In Section 4, we provide the simulation results for our model. Section 5 offers our conclusions and prospects for future research.

1. Vehicular Communication Networks and Their Architecture

Vehicular communication networks can be formed spontaneously between the moving nodes that are equipped with the homogeneous or heterogeneous wireless interfaces (802.11a/b/g/n/p, WiMax, 3G, LTE and so on.). These networks, also known as the VANET (Vehicular Ad-Hoc Network) is one of the MANET (mobile ad-hoc network) applications, allowing communication between the nearby vehicles and vehicles and stationary equipment (road side units). Vehicular communication application areas can be divided into three main categories: general information - multimedia services, road safety and traffic monitoring and management services [2].

An analysis of the scenarios where the communication is made between the sender and the recipient moving in the opposite directions was made; it is given in Table 1.

Table 1. Scenario analysis of the vehicular communication network

	Rural	Town	City	Highway
Average speed of the nodes	Average	Low	Very low	Very high
Node density	Low	Average	Very high	Average/low
Interference	Low	Average	Very high	Low

1.1. The Specific Characteristics of the Vehicular Communication Networks

Vehicular communication networks have special characteristics and properties that distinguish them from other types of mobile communication networks. According to [3] and [4], it was summarized the following unique features:

- High energy reserve;
- Huge mass and size of the vehicle;
- Moving by the patterns.

Vehicles have much greater energy reserves, compared with a conventional mobile device. Energy can be obtained from the rechargeable battery and gasoline, diesel or alternative-fuel motor. The vehicles are many times greater and larger compared to traditional wireless devices, and therefore, can support a much greater and heavier computing, radio and sensor components. Computers can be bigger, faster, and provide very high-capacity memory devices (terabytes of data), and powerful wireless interfaces, capable of high speed communication. The vehicles can move at very high speed (160 km/h or more), making it difficult to maintain a consistent, coherent V2V communication. However, the existing statistical data on vehicle movements, such as the movement together according to certain patterns or peak time can help to maintain a link between the mobile automotive groups. Vehicle at any time may be out of communication coverage (WiFi, cellular, satellite, etc.), so the network protocols must be designed so that it can easily connect to the Internet, in normal mode. Despite the many positive unique features, vehicular network's development is faced with specific challenges, as their primary:

- Large-scale networks;
- High level of mobility;
- Fragmentation of the network;
- Changing topology;
- Complex communication quality assurance.

Unlike the literature described ad-hoc networks, which are quite limited in size, vehicular communication networks, in principle, can extend across the road network and cover a huge amount of network equipment (vehicles). The environment in which networks are operating is extremely dynamic and, in some cases it may be highly different, for example, in highway speeds can reach up to 300 km/h, in the low-density roads car density may be as only about 1-2 cars kilometer. On the other hand, the speed of cars in urban areas is 50-60 km/h and the car density is quite high, particularly during the peak periods. Often vehicular communication networks may be fragmented.

The dynamic nature of traffic can lead to large gaps between cars in sparsely populated areas; it can also be created a few isolated clusters of network nodes. Vehicular communication networks' scenarios are highly different from the classic ad-hoc networks, since the cars are moving and constantly changing positions, scenarios are highly dynamic. Furthermore, the network topology changes extremely frequently, since the very frequent connections and disconnects between network nodes. In fact, the degree to which the network is combined depends on two factors: the distance between the wireless nodes and number of connected vehicles [5].

2. Related Work

There is a growing literature on data-transfer capabilities for heterogeneous service support within vehicular networks, some of which have also considered the application of our analyzing problem. We briefly discuss the key the most recent relevant references next, and highlight their difference from our approach.

Analysis of the performance of DSRC-based VANETs in delivering CVSS (Cooperative vehicle safety systems) messages was made in [6]. Here a network performance measure is defined, which can be used as an indicator for the success of

CVSS tracking application. A study into how controllable parameters such as rate and range of transmission affect this performance measure has revealed interesting properties of IDR. It is shown that robust control of rate or range of transmission based on the relationship between IDR and channel occupancy is possible. Based on these concepts, a robust range control method is analyzed and evaluated.

Another performance evaluation of information propagation in a vehicular ad-hoc network was made in [7]. The authors' studies packet loss rate, expected transmission distance and effective coverage range of road-side station. They state that communication performances are similar under three distributions in most cases where negative-exponential distribution shows the worst performance. It can be assumed that under negative-exponential distribution, the randomness of space headway is strong, this will break down the connectivity of the communication chain.

In [8] presents results for 35 field trial data sets collected in Australia, Italy, Germany, Austria, and the United States, covering over 1100 km on the road in a wide variety of physical environments. The performance results reveal that DSRC/ WAVE can provide highly reliable communications, and sufficient driver warning times in support of the targeted road safety applications. However, analysis of channel sounding data collected shows that NLOS safety-critical conditions require careful attention to physical layer receiver processing in order to provide a safety benefit.

The performance modeling of message dissemination in vehicular ad-hoc networks with two priority classes of traffic was presented in [9]. The results showed that the probability of a receiving node being exposed to interference increases as a function of the transmission range, and that this increase is faster at higher-density node traffic.

Performance of the 802.11p Physical Layer was estimating in the [10]. Authors have found that the primary problem is that the channel estimation mechanisms built into the 802.11p standard only allows for channel estimation at the beginning of each packet. Because the packet length is not restricted by the standard, the initial channel estimate can expire before the packet has completed transmission. They state that, the channel estimate must be updated throughout the length of the packet. Furthermore, authors make the conclusion that the maximization of throughput is a tradeoff between high overhead at short packet lengths and poor performance at longer packet lengths.

Nevertheless, the growing number in of researches in terms of data-transfer capabilities in vehicular communication networks, none of them investigates a special scenario where the nodes are moving in the opposite direction in a highway. From this point of view, our work is different and novel.

3. Methodology and Experimental Model

As it was mentioned in the previous section, the services in the vehicular communication networks can be classified into the road safety, information and multimedia services' categories. To support high-quality services it must be taken into account the data rate, packet delivery efficiency and collision rate. The analysis shows systematic data quality requirements for different services for vehicular communication networks (Table 2). To determine the influence of the number of vehicles in connection capacity it was made a number of experiments which goal is to evaluate data-transfer efficiency when providing mobile multimedia services in the communicative network between in the opposite directions moving sender and receiver nodes at high speed.

Table 2. Data transmission quality requirements for different services support in vehicular communication networks, by the [11, 12]

Service	Packet size (in bytes) / required throughput (KB/s)	Packet loss influence	Periodicity of transmitted data	Tolerated latency (ms)
Road safety services				
Lane changing	~100 / 1	Average	Event	~100
Traffic light control	~100 / 1	Average	Periodic	~100
Warnings about dangers	~100 / 1	High	Event	~100
Warnings on road conditions	~100 / 1	Average	Periodic	~100
Multimedia services				
IPTV	~1300 / 500	Average	Periodic	<200
VOIP	~100 / 64	Average	Periodic	<150
Video/audio files exchange	As high as possible	High	Periodic	-
Games	As high as possible	High	Periodic	-

The experiments were carried out in the simulation environment NCTUns 6.0 [13], which was installed on Fedora 12 Linux operating system. The environment was chosen as it uses the existent Linux TCP/UDP/IP protocols stack, it provides high-accuracy results; it can be used with any actual Unix application on a simulated node without additional modifications; it supports 802.11a/b/p, 802.16e communication networks and vehicle mobility modeling, user-friendly user interface, and it is capable of repeated the simulation results. In the experimental scenario (Figure 1), a node (4) sends data to the node (11). Communication is provided via 801.11b standard interface and is used multi-hop data transmission method.

It was analyzed and structured requirements for the NCTUns simulation model (Table 3). The experiment was carried out when the number of nodes in the network is from 10 to 100 - simulating different traffic congestion to determine the impact of the vehicle's number for the data-transfer efficiency. Senders and receiver's nodes are moving at high speed (130 km/h) in the opposite directions. The remaining vehicles are moving at different speeds from 90 km/h to 150 km/h, and their speed and directions of movement are spread evenly. These parameters are chosen to simulate the realistic movement of cars on highway conditions.

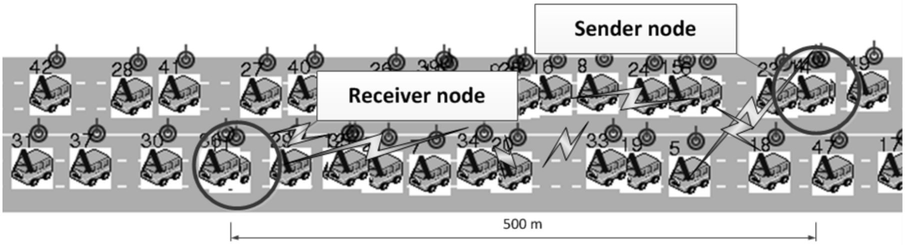


Figure 1. The experimental scenario

Table 3. Simulation parameters for the experiment

Parameter	Value
Simulation time	60 s
Physical layer protocol	802.11b
Number of nodes	from 10 to 100
Nodes mobility model	Random, highway
Channel frequency	2,4 GHz
Routing protocol	AODV

4. Experimental Results

During the experiments, it was evaluated data transmission efficiency – outgoing throughput, download throughput, packet drops and collisions with a different number of vehicles on the network. The data was transmitted using the UDP protocol, and a packet size of 1000 bytes. Simulation was carried out for 60 seconds. The assumption was made that the communication time between the sender and the recipient is directly proportional to the number of cars on the network. Furthermore, with increasing number of nodes it is expected to increase the collision rate and rejected packets.

Analysis of the data collected during the experiments shows the download speed versus time, with a different node's number on the network (Figure 2). The graph shows that the longest communication time is achieved by operating the largest network of vehicles - 100. With the maximum number of vehicles, the network coverage increases, so the data can be transferred for a longer period of time. With 100 vehicles and about 330 KB/s data transfer rate, we have managed to maintain communication for 30 seconds. The speed from 31 s decreased to 50 Kb/s, but from 37 s to 41 s the rate rises to 230 Kb/s, and from 46 s to 48 s - to 130 KB/s. When the vehicles passed each other the connection was lost. The minimum data rate was achieved by the network operating 50 vehicles. Moreover, in this case, the shortest communication time is achieved. With a small number of vehicles (10-30), it is maintained a relatively high data transfer rate, due to the low collision rate.

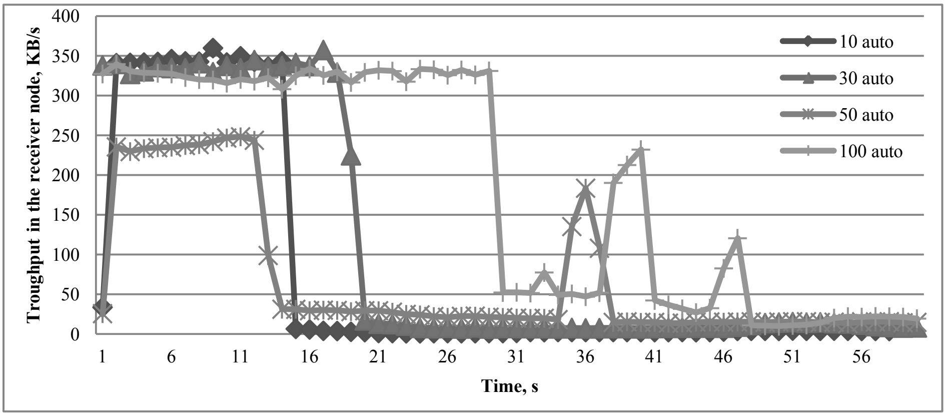


Figure 2. Data download rate dependence from time with a different number of vehicles in the network

After the experiment, the other important parameter - the average data uplink and downlink throughput was measured (Figure 3). In this case, the highest mean transfer rate achieved by the network operating 20 vehicles, while the meanest - 30. The maximum average data rate of downlink – 100 vehicles, while the meanest – 50.

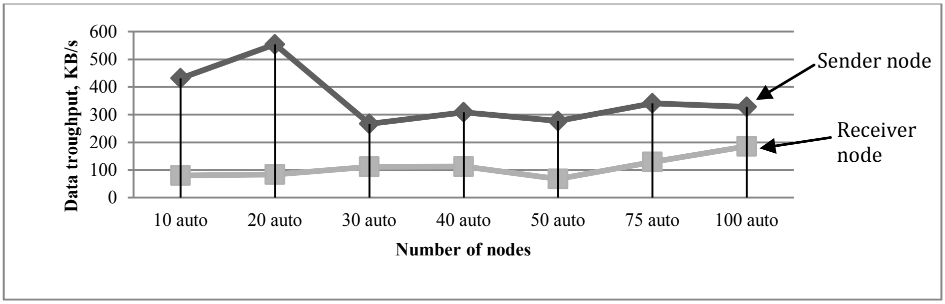


Figure 3. The average data downlink and uplink throughput with a different number of vehicles

It was found out collision's dependence on sender and receiver nodes with a different number of vehicles (Figure 4). Collision rate is directly proportional to the number of vehicles. Up to 40 vehicles, collisions rate at the receiver and sender nodes is similar, but from 50 vehicles, collision is greater in sender node because of unsuitable channel access mechanisms.

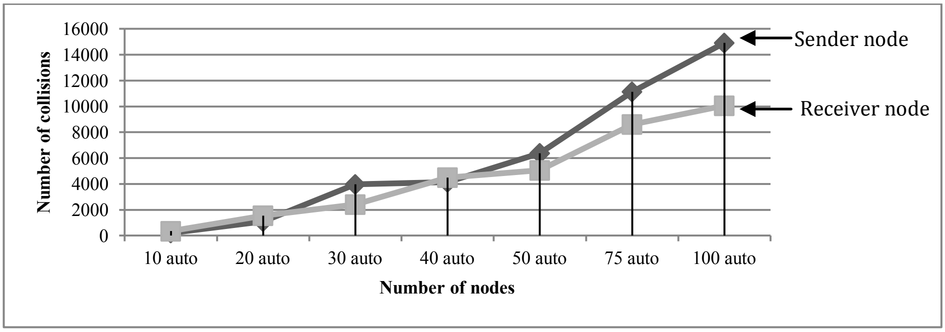


Figure 4. Collisions rate dependence on receiver and sender nodes with a different number of vehicles

5. Conclusions

It was performed the experiments in which was investigated communication and data-transfer efficiency between at high speed moving sender and receiver nodes in the opposite directions, in the mobile multimedia services communicative network. The goal was reached, and it was estimated the transmission efficiency and quality of communication. It was found that the longest communication can be maintained at the maximum number of vehicles, but that communication quality is inversely proportional with the number of vehicles, as the increasing number of vehicles - increasing data and network flooding occurs in many collisions.

To provide quality heterogeneous services it is necessary new routing protocols and channel access methods for the large volume fast changing topology networks. This investigation is important because it examining problems associated with communication between the sender and the receiver moving in opposite directions in highway, where the network topology varies very rapidly and, which may contain from one to several hundred of the network nodes. Future plans to extend the study to include other proactive, reactive and hybrid (ADV, DSDV, AORP, etc.) routing protocols.

Acknowledgements

The authors thank the Projects LLII-061 Development of Joint Research and Training Centre in High Technology Area and LLIV-215 JRTC Extension in Area of Development of Distributed Real-Time Signal Processing and Control Systems for the possibility to complete a scientific research.

References

[1] H. T. Cheng, H. Shan, and W. Zhuang, Infotainment and road safety service support in vehicular networking: from a communication perspective, *Mechanical Systems and Signal Processing* **25**(6) (2011), 2020-2038.

[2] T. Willke, P. Tientrakool, and N. Maxemchuk, A survey of inter-vehicle communication protocols and their applications, *IEEE Communications Surveys & Tutorials* **11**(2) (2009), 3-20.

[3] H. Moustafa and Y. Zhang, *Vehicular Networks : Techniques, Standards, and Applications*, CRC Press, Boca Raton, 2009.

[4] U. Lee and M. Gerla, A survey of urban vehicular sensing platforms, *Computer Networks* **54**(4) (2010), 527-544.

[5] G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, and T. Weil, Vehicular networking: a survey and tutorial on requirements, architectures, challenges, standards and solutions, *IEEE Communications Surveys & Tutorials* **13**(4) (2011), 584-616.

[6] Y. P. Fallah, C.-L. Huang, R. Sengupta, and H. Krishnan, Analysis of information dissemination in vehicular ad-hoc networks with application to cooperative vehicle safety systems, *IEEE Transactions on Vehicular Technology* **60**(1) (2011), 233-247.

[7] Q. Wang, J. Hu, and J. Zhang, Performance evaluation of information propagation in vehicular ad hoc network, *IET Intelligent Transport Systems* **6**(2) (2012), 187-196.

[8] P. Alexander, D. Haley, and A. Grant, Cooperative intelligent transport systems: 5.9-GHz field trials, *Proceedings of the IEEE* **99**(1) (2011), 1213-1235.

[9] M. Khabazian, S. Aissa, and M. Mehmet-Ali, Performance modeling of message dissemination in vehicular ad hoc networks with priority, *IEEE Journal on Selected Areas in Communications* **29**(1) (2011), 61-71.

[10] J. A. Fernandez, K. Borries, L. Cheng, B. V. K. V. Kumar, D. D. Stancil, and F. Bai, Performance of the 802.11p physical layer in vehicle-to-vehicle environments, *IEEE Transactions on Vehicular Technology* **61**(1) (2012), 3-14.

[11] C. Sommer, A. Schmidt, Y. Chen, R. German, W. Koch, and F. Dressler, On the feasibility of UMTS-based traffic information systems, *Ad Hoc Networks* **8**(5) (2010), 506-517.

[12] F. M. V. Ramos, J. Crowcroft, R. J . Gibbens, P. Rodriguez, and I. H. White, Reducing channel change delay in IPTV by predictive pre-joining of TV channels, *Signal Processing: Image Communication* **26**(6) (2011), 400-412.

[13] S.Y. Wang and C.L. Chou, NCTUns tool for wireless vehicular communication network researches, *Simulation Modelling Practice and Theory* **17**(7) (2009), 1211-1226.

Learner Model's Utilization in the e-Learning Environments

Vija VAGALE and Laila NIEDRITE
Faculty of Computing, University of Latvia, Raina boulv. 19, Riga, Latvia
vija.vagale@du.lv, laila.niedrite@lu.lv

Abstract. In the field of personalized systems big role is granted to the adaptive e-learning environments. The task of these systems is very important and complicated. With their participation the learner gains exactly the knowledge he needs most, and the system adapts to user needs, expectations and his individual features. In this kind of systems information about learner is saved in the learner model also known as the user model and student model. For the system to be able to perceive and analyze user activities correctly, is necessary to define what kind of information about the learner has to be saved. The article gives an overview about already existing learner models, their utilization methods and also it offers their comparison according to various criteria.

Keywords. E-learning, adaptation, learner model

Introduction

When the tempo of life becomes more and more intensive, the necessity of the new and effective solutions in different scopes including education arises. One of the most actual tasks for educational quality improvement is the utilization of an e-learning system. One part of such learning environments is passive and used only to supply users with static content and to ensure identical system reactions to the users' activities. The other part of the learning environments adapts to the user as a personality by offering learning in the most appropriate way for him. The abovementioned environments are called adaptive learning environments (ALE). One of their tasks is to find out user's personal qualities that influence his learning process and his knowledge level in certain moment of time to offer a learner certain learning content and learning methods, which are the most appropriate exactly for him, to ensure the best learning results. Several adaptive systems are offered in the works of Hauger and Köck [14].

The role of adaptive systems nowadays gets bigger every year. This type of systems can help to acquire knowledge at schools, universities and other kind of learning institutions. ALEs can be used for the student teaching in schools as a secondary instrument to acquire fundamental and additional knowledge, but for students in universities they serve as a primary instrument for acquiring new knowledge and organizing their study plan. For the lifelong learning of adults ALE can serve as an instrument for gaining new knowledge and raising professional qualification. ALE can be used also to ensure learning interaction with already existing learning environments. For example, ALE cooperating with popular social networks can offer some specific knowledge for certain user group. Also, for the children of preschool age adaptive learning system can be used as an instrument of gaining basic knowledge that would be necessary at school.

The aim of the work is to explore structure models of the adaptive systems (including learning systems), especially – user model and its components, and to explore data which are already included in user model and select the most common used data which would be useful for creating adaptive system based on user model.

This work is written based on scientific articles that reflect the newest trends in formatting and utilization of the learner model. The most-cited literature has been analyzed, as well as the newest review articles about learner model and papers that describe already realized adaptive system examples.

Alenka Kavcic [16] points out several most relevant questions, which should be considered when creating a user model: a) what kind of information must be included in the user model, b) how to gain this information, c) how to represent information about a user in the system, and d) how to create and restore a user model.

The first section of this paper gives an overview of the models used in the adaptation process, user profiles and explanation of the learner model concept, their common and different features, the ways to obtain profile data and its utilization for creating a user model. In the second section an overview of the data included into the user model of an adaptive learning system, and the data included into the learner model are summarized. Also, a definition of this data is given. In the end of the section LM is offered a table that summarizes employed data with the aim to obtain the most significant data that must be included into learner model. The third section describes the creating stages and construction techniques of the user model of an adaptive system. The paper ends with conclusions on the accomplished overview of the learner model utilization in adaptive systems.

1. Learner Model Essence

1.1. A Review of the Models Necessary for Adaptation

Adaptation in the learning environments is based in the well-organized models and processes. Data that describes knowledge in the system and learner behavior are extensive. When exploring scientific articles [8, 9, 16, 24, 28, 33] on different types of adaptive systems, one may conclude that they are based on the three main models: a domain model, a user model and an adaptive model.

A domain model includes two main parts: content or system offered domain knowledge, and a supply system for this content. In [9] authors points out that a domain model works like a data repository, which consist of topics, content, pages or nodes and navigation links that connect the represented data design structure. Domain knowledge consists of knowledge basic elements such as concepts, topics, knowledge items, learning goals, learning results. Domain delivery system must support all course types and manage to adapt to the different requirements for the course content.

An adaptive system adapts to users' needs that is why one more important component of these systems is a user model. In the learning systems it is also called a student model or *learner model* (LM). In [26] authors call a LM the key and core of the adaptive system. The learner model keeps all information about the learner, i. e., personal information, his knowledge, skills, and behavior in the system. Intelligent Tutoring Systems (ITS) learner model is also called a student model [35].

Domain and learner models are connected with the help of an adaptive model. *An adaptive model* ensures the application of the system flexibility theory by combining domain and learner models [9]. By analyzing learner model student needs are gained, and knowledge representative nodes are offered to him the system. These nodes can be classified by knowledge type as follows: basic knowledge (includes knowledge about definitions, formulas, etc.), procedural knowledge (solves relationship between stages), and conceptual knowledge (refers to relationship between concepts by developing bigger common scene) [39]. In widely spread Adaptive Hypermedia Systems (AHS) an adaptive model is also called an interaction model [11, 27]. In the Intelligent Tutoring Systems (ITS) adaptive model functions are fulfilled by pedagogical model [35].

Depending on the adaptive system type, in addition to the previously mentioned models, there can also be other models, which ensure the system supplied services. AHS systems have also the fourth model: media model [3, 28]. Along with the learner model there is also a group model [33], which is similar to learner model but is filled dynamically and is based on learner group identification after some common features and behavior.

1.2. Basics of Learner Model Creation – A User Profile

To make an adaptive system, which could respond exactly how the user wants, information about the user is needed. The easiest way how to obtain information about the system user is to use his data from the user profile.

In the profile static (constant) information about the user without any additional description or interpretation is kept [30]. Profile data contains learner personal data as well as data on his individual features and habits. User data in the profile is represented as attribute pairs – key-value. User model creation, modification and maintenance process is called user profiling. A system should provide profile attribute initialization, adding, saving, modification, deletion and extraction.

Unlike the profile, a *user model* is an abstract representation of the system user [23], where, in addition to the profile data, some specific information about the person is included. For example, in [27] a learner model consists of the domain independent data that consists of Generic profile, Psychological profile, and domain-dependent data. The user model contains all information that the system has on the user and maintains live user accounts in the system [33]. In the general case, the profile concept is narrower than the user model concept. In a simplified case, they can coincide [Kules2000]. User profile data can serve as the base for the creation of the user model. A profile keeps static information, but the user model keeps both – static and dynamic information.

1.3. Obtaining Data for the Learner Model

When the user interacts with the system for the first time, a user profile that contains the basic information about the user is being created in the system. Information in the user profile can be obtained in similar ways.

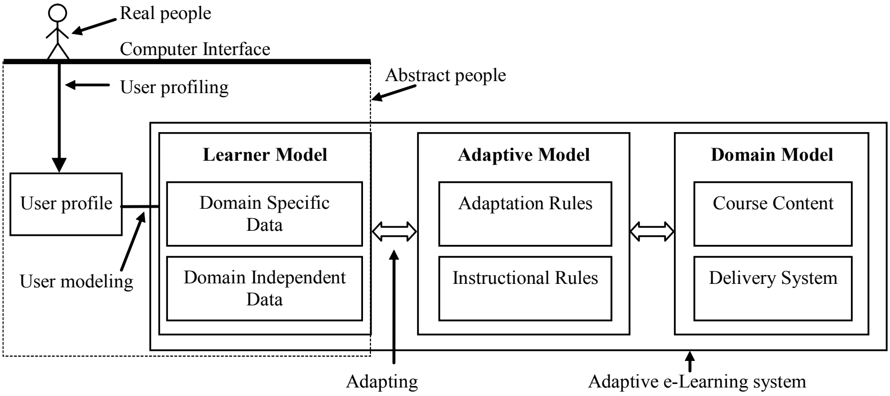


Fig. 1. Adaptive e-Learning system scheme

There are several approaches to create a user profile:

a) A user creates his profile on his own, based on his interests [25]. A part of the user profile information can be obtained directly from the user registration form or questionnaires: for instance, birth date and gender. For example, in [7] EASEL project with the help of AHS Questionnaire Servlet the system asks the user some simple questions to gain data about visual, audio, read/write and kinesthetic attributes. However, such information as user preferences is very hard to gain and that is why they must be taken from the user interaction with the system.

b) A system creates a profile by itself by collecting necessary information about the user indirectly, for instance, from activity log files, where it is written what a user has chosen and what actions he has accomplished in the system [25].

c) Mixed approach, when one part of information is input by the user, but the second part of the information the system gains indirectly [25].

d) ALE integration with other informational systems with user data import from some other system:

- From the informational system (IS) which ALE is related to. For example, ALE has "cooperation" with administrative system of the educational institution, which contains general information about an IS user. IS imports user data into the adaptive learning environment and ALE uses this data as an entrance data to create the first notion about the learner. When gaining data from IS to the ALE in this way, such information as the type of knowledge student must acquire (for example, registration to the certain course) is also often indicated [35].
- From other type of systems with user registration. Nowadays social networks where people spend a lot of time (for example, facebook.com, draugiem.lv) have become widely used and popular. Versatile information about the user is saved there, i.e., his personal data, interests, skills of communication with people, activities in groups, etc. From that kind of systems, which widely characterize its user, data can be taken and integrated into united adaptive learning system user model. In this case, it would be important to anticipate both system cooperation opportunities.

e) Gaining data from ePortfolio: a web-based electronic material resource, which contains material collection that is made and managed by user [5, 42]. ePortfolio also indicates the user learning or professional growth. Fsor example, in [34] a research about ePortolio integration with Learning Management System (LMS) Moodle is described.

2. Learner Model Data

2.1. Learner Model Data Types

Information included in the user model can be grouped differently: (a) by data dependence from the subject, (b) by data obtaining type, (c) by data availability for the learner, and (d) by data life-cycle in the user model.

Relative-to-subject information can be *"domain-dependent"* and *"domain-independent"*. Domain-dependent information shows the knowledge level and ability of the learner at the certain moment of time. Domain-independent information is not dependent on the offered content; it is for example, motivation, skills, learning style.

In educational AHS (Brusilovsky [4]) LM data is divided into two big groups: *domain-specific information* (DSI) and *domain-independent information* (DII). Domain-specific information contains student knowledge model that describes student knowledge level, insight about knowledge, learner mistakes and records about learning habits and ratings. The domain-independent information includes information about the learner skills based on his behavior. DII includes also learner learning goals, his cognitive abilities, motivation, background, experience and preferences. The work [27] also has similar data included in the user model division: *domain-independent data* (DID) and *domain-dependent data* (DDD). Domain-dependent data stores specific learner knowledge from the domain that system concludes about the learner. Domain-independent data includes two elements: the Psychological Model and the Generic Model of the Student Profile [22]. Psychological data are connected with the student exploration and emotional aspect. The Generic Model of the Student Profile keeps user interests, common knowledge and experience.

Another LM division is based on the way how data about the learner is gained: *"content-based"* or *"collaborative"*. In content-based case, data are collected or concluded about the learner only from his interaction with the system. In collaborative case, data is obtained from other similar learner groups that share, for instance, similar interests and necessities [40], and is used for a certain learner.

The user model data can be *"visible"* and *"opaque"*. Visible data can be changed by user with the help of questions-answers offered by the system. On the contrary, opaque LM data are not available for the user [19].

In [25], [26], [40] the division of data by its life-cycle in the system used in the learner model or by learner interaction with the system (values are changing or not):

- *Static data* is data that are not changed during the student and system interaction;
- *Dynamic data* is data that changes depending on the student learning progress and interaction with the system.

Static data types are personal, personality, cognitive, pedagogical and preference data [13]. Static data is collected either once at the beginning of system utilization or after a determined period. This data stays unchanged during the system utilization. Dynamic data is gained based on learner interaction with system. Dynamical data is divided into *performance* data and *student knowledge* data. *Performance data* is data gathered from the user and system interaction, and this data summarizes the information about student's achievements during the course session and is being continuously restored. *Student knowledge data* is data that describes knowledge concepts and competences. This data set gathers information about student progress in the course of learning.

2.2. Analysis of the Data Included in the Learner Model

In this section eight works describing adaptive systems have been considered, and data included into user model have been analyzed. The research was complicated because in data categories used in these works were similar by meaning but with different names and vice versa, had similar names but category content was different. By making article research the most important data categories used in viewed scientific works were selected: personal data, pedagogical data, preference data, personality data, cognitive data, history data, device, context, interests of user, interests gathered by system, results of assesement, domain expertise, acquired knowledge, performance data, deadline extend, student knowledge currently. Below one may observe the categories and data included in them depending on the category occurrence in the viewed works.

Personal data is a category without which no system can do. It includes user biographical information gained, for instance, from the registration form. This category combines personal information, demographic information and identity data. Data included into the personal data category is similar for all authors:

- In [13] authors include in this category student name, special accessibility needs to course materials that the student must have; affiliation; student's professional activities; list of degrees and qualifications; information about student security and access credentials.
- In [10] personal data is gender; age; language; culture; name; email; password.
- In [27] personal information is name; email; password; demographic information is age.
- In [1] personal data category includes name; surname; age; language; media type; login.
- In [41] as demographic information: gender; age; native language; social-cultural parameters: formal education, family income is considered as personal data.
- In [25] authors include in personal data category personal information: name; age; address books; demographic information: date of birth; gender; nationality.

Next category that is mentioned in almost every work is *pedagogical data* that describes how and what to learn. This category includes programs, topics, course collections and course sequence. In the majority of examined scientific works this category includes also learner knowledge before adaptive system utilization, for some authors this data was described in the personal data category. Data sets included in the pedagogical category of each article author significantly differs, for instance:

- In [13] this category consist of learning style; learning approach; course objectives – concept list that each student must acquire during the course session; course evaluation; course navigation control, i.e., how the course content is navigated.
- In [27] authors include in this category academics backgrounds, for example, technological studies contrary to economically; qualifications: certificates;

background knowledge: knowledge collection that is transformed into concepts; ability to determine qualitative, quantitative and probably user acquired concepts and knowledge; background knowledge: knowledge collection translated into concepts; plan.

- In [1] student knowledge level in certain concepts of the onthology-based learner model is marked with very low, low, medium, good or excellent.
- In [3] Brusilovsky calls this data a background that contains basic knowledge with which the user started to employ the adaptive system.
- In [29] authors include skills, knowledge, abilities and plan in this category.
- In [41] the pedagogical data category includes the learning plan.

In more than a half of scientific works data that falls into the category of *preference data* is observed. This category keeps student preferences relating to the system adaption. Majority of the preferences are taken from the student, but the remaining part is defined by system administration. Characteristic examples of this category are the following:

- In [13] this category includes preferred presentation format; preferred language for content display; web-design personalization; command personalization; personal notebook; sound volume; video speed; subtitles.
- In [27] authors include in the preference data category learner defects, for instance, bad sight; domain of application user localization.
- In [1] authors include in this category specific data that is necessary for organizing learning process based on learner ontology model.
- In [3] Brusilovsky explains that the system cannot calculate this data, therefore, it is gathered directly or indirectly based on the user activities.

The next category that was distinguished is *personality data*. This category gathers data that describes student as a personality: learning style, concentration skills, collective work skills, relationship creating skills, individual features and attitude towards learning. A part of these data can be gained via tests. All in all, authors in this category have included similar data, for example:

- In [13] authors combine in this category personality type; concentration skills, where the base is average time that is used for learning; collaborative work that characterizes student skills to work in groups; relational skills that characterize student skills to communicate with teacher.
- In [27] authors include in personality data category learning style; information reception abilities – cognitive capacities; traits of personality: introvert, extravert; activity; inheritance of characteristics that classifies users, so that further system could create learner stereotype models.
- In [41] this category includes specific data – emotional state, it is ability of adaptive system to model determined user emotions with purpose to make system behavior correction.
- In [29] authors include attitude in personality data category.

For successful adaptive system utilization it is necessary to have data on user experience in work with computer, certain software and adaptive system. This data is described by the half of all authors, the data type are gathered in the category *system experience*, for instance:

- In [13] authors keep in the user model experience level that describes student's ability to work with an e-Learning system and student's experience in computer utilization.

- In [3] Brusilovsky defines experience as how familiar the user is with system and how easily he orients in it.
- In [27] this category is called aptitude.

In the half of the reviewed scientific works the category *goal* or *motivation* is presented. This is data about the system user long-term interests [25]; data that characterizes reasons why the learner does some actions (for example, searches and uses specific information) [10].

Cognitive data category describes what reference types a student has. These characteristics can be obtained by tests. This data has an important role in the system adaption ability to the learner. These data types are found only in three of eight reviewed works. For example, in [13] this category includes cognitive styles. Authors in [10] describe the data that characterizes the way of how the user processes the information.

Below are listed categories which include data used in several works only. *History data* category includes data that contains information about user activities in the system. For example in [10] information about the last user interaction with system (log file) is considered. In [27] authors use data about access of each page. In [1] authors include in history data category the data on learner navigation during resource learning process.

Device category incorporates data that characterizes user environment during the adaptive learning system utilization. In [10] it is hardware; screen size; download speed. In [27] this is data connected with the user environment, for instance, screen resolution.

Context category incorporates data which characterizes the user access place. It is relevant in cases when someone is using different devices to access the adaptive system. This data intersects a little with the device category. In [10] in this category authors keep information about the access environment, for example, access from home or from educational institution. In [41] wider information is included in this category: user location, time, physical and social environment, used devices, etc.

In the viewed papers Interests category is mentioned. For one part of authors those are interests that the user indicates himself, for the other – interests collected by the system based on the user activities. For better understanding of the observed interests, this category was divided into two parts: *Interests of user* and *Interests gathered by system*. In [10] the system collects interests from user keywords and searching results. In [27] authors examine person's interests, which are used to adapt navigation and content.

For saving learner knowledge data categories that characterize specific knowledge types are used. *Results of assessment* category contains the learner knowledge test data, for example, [27] keeps data about all tests and exercises. *Domain expertise* determines knowledge in topics that the user has interest for [10]. *Knowledge acquired* describes learner knowledge in the certain moment, for example, in [27] it is mentioned that the learner knowledge is transformed into concepts.

Summary about the examined data categories of user model is shown in Table 1. Table column names correspond to the researched articles, and rows – to data categories. Category utilization in certain article is represented with a "+".

When analyzing the results of Table 1, it is obvious that personal and pedagogical categories are the most common in the learner model. In the works of several authors there is no precise borderline between categories personal and personality and the same data is included into personal and personality categories. Cognitive data is similar to personal and personality data, where data that characterizes learner and significantly affects learning process is saved. One of the widest and most important categories of the user model is pedagogical category, which incorporates learner basic knowledge and learning plans.

Table 1. The frequency of the learner model data category utilization in scientific works

Data category type	[29]	[3]	[25]	[41]	[13]	[10]	[27]	[1]
Personal data	+		+	+	+	+	+	+
Personality data	+			+	+		+	
Cognitive data/style					+	+	+	
Pedagogical data	+	+	+	+	+		+	+
Preference data	+	+			+		+	+
History						+	+	+
Device				+		+	+	
Context/Environment	+			+		+		
Interests of user	+			+			+	
Interests gathered by system						+	+	
Goal/Motivation		+	+	+		+		
System Experience		+			+	+	+	
Domain Expertise						+		
Results of assesment	+				+		+	
Knowledge acquired	+						+	
Deadline extend							+	

3. Learner Model Modeling

An adaptive system continuously collects data about the learner. This process is called user modeling and it is quite complicated. In this process the activities in LM, mechanisms used in modeling LM and the way in which LM saves data must be taken into account.

3.1. Formation Stages of the Learner Model

First, the adaptive system initializes the user model, and only after that data is being refreshed in the LM. Several authors highlight one more stage that involves learner data mining and concluding. Extensive research about the educational data mining is found in the review article [38].

After making article analysis, the authors of this paper concluded that the most widespread l*earning model formation stages* are the following*:*

- Initialization – information and data gathering about the user, and user profile formation that is based on obtained information. During the initialization process the structure of the user model is defined as well as reasoning methods, and memory (i.e., information about user abstraction state in a certain moment of time) in the user model. There are two ways how to obtain data: explicit questions and initial tests. This stage is used in the articles [9, 16, 25, 30].
- Updating – the system must be ensured with the learner model actualization; the system observes user activities, evaluates user achievements, learning progress dynamics and makes the reflexive link analysis from the user's

interaction with the system. All these activities are executed by the system implicitly or explicitly. This stage is covered in the articles [9, 16, 25, 30].

- Reasoning (i.e., extraction of the new information about the user from the existing available data) [13]. Data mining utilization to ensure adaptation is a new research direction which is reviewed in [21, 38].

3.2. LM Construction Techniques

The user modeling technique describes how the user model is created and maintained. There are many techniques for modeling the user model and supplementing data in it. For example, there are such LM construction techniques as: a stereotype model, an overlay model, a combination model, a perturbation model, a plan model. The oldest and the most frequently used ones are a stereotype model, an overlay model and a combination model [6].

A Stereotype model is often used for LM to define some default values. In case of the learning system, users are divided in the system-offered categories, i.e., stereotypes. It has been introduced by Elaine Rich. A stereotype is a simple collection of the aspect–values that describes the system user groups [36]. The benefit of it is that with a small amount of information it is possible to conclude a lot of new assumptions about the user. The stereotype review and utilization examples are given in articles [3, 6, 7, 12, 15, 20, 30, 37].

An Overlay model is widely used in adaptive hypermedia systems. A student knowledge model is being constructed based on the concepts. A user model is restored based on the user progress in the system. This model allows creating knowledge about a student in each topic in a flexible way. The modeling approach of the overlay model is based on the domain model, which is often constructed with the help of the knowledge network or the knowledge hierarchy tree. This method requires dividing domain model in different topics and concepts [3]. Nowadays domain model can be built with the help of ontology [30]. The overlay model review and utilization examples are provided in articles [6, 7, 15, 27, 30].

A Combination model employs both of the previously mentioned models. First of all, students are divided by stereotype, and then this model is gradually modified into an overlay model. This model is used for educational AHS [6].

A Differential model is described in [30]. It is another version of overlay model, where the knowledge that a student must acquire in a certain period of time is represented (i. e., expected knowledge). This knowledge can be considered as the knowledge that is missing.

A Perturbation model. Some models are not interested in learner mistakes caused by wrong perception or lack of knowledge. This model represents learner knowledge as an overlay model plus his mal-knowledge [30].

A Plan model incorporates successive student actions for achieving certain goals and desires [30].

3.3. User Data Modeling Methods

Static data elements are modeled with *Attribute-Value Pairs* [25]. Attributes are terms, concepts, variables and facts that are important for both the system and the user. Their values can be of the following types: boolean, real or string. For modeling uncertain information elements like user knowledge more difficult approaches there are used such as rules with certainty factors, fuzzy logic, Bayer probability networks or Dempster-Shafer theory of evidence [16]. User condition-based language modeling approach is applied to the *dynamical data* elements, where a connection between the provided service and a context is based on if-then logic.

To represent the relationship between data elements the hierarchy tree modeling approach and ontology are used. The user ontology is described in following articles: [1, 13, 29]. In [43] a research about e-learning platform knowledge management with the help of ontology is described. Systems used for user modeling are:

- UMT [2] that allows developers to define hierarchically arranged user stereotypes, rules of the user model conclusions and contradictions.
- PROTUM [44] represents the user model content as a list of constants, where each constant is attached to a certain type and confidence factor. This tool has deeper stereotype retraction mechanisms than UMT.
- TAGUS [32] represents assumptions about the user with first-order formulas by indicating different types of assumptions. This tool allows defining stereotype hierarchies and contains conclusion mechanisms.
- UM [17, 18, 19] toolkit3 includes user modeling by indicating user knowledge, views, desires and other user characteristics with attribute-value pairs.
- BGM-MS [19] user or user group stereotype choice is based on the assumptions that are gained using predicate logic. User knowledge defining is based on conclusions that are obtained employing different assumptions.
- DOPPELGÄGER [31] is a server that collects information about the user with hardware and sensor software. A user can visualize, check and edit his data.

4. Conclusions

Adaptive system range is wide and there are a lot of researches in this scope including the user model utilization.

During the analysis of the adaptive system structure authors concluded the following: (1) depending on the type of the adaptive system, model names included in it may differ but their essence and tasks remain similar; (2) each adaptive learning system must have at least three components: (a) a domain model for keeping system-offered knowledge; (b) a learner model (user model, student model) which describes in an understandable way for the system a person sitting in front of the computer and willing to learn; (c) an adaptive model (interaction model) with the help of which system-offered knowledge is delivered to the learner in an understandable way.

One of the most important adaptive learning system components is the learner model. It includes data from the user profile, however, when making a good adaptive system learner model can also include additional data characterizes the learner in a more comprehensive way.

It would be recommended to divide all data included in the learner model into some basic categories:

- Personal data, where data about the personality identity is stored (name, surname, login, password, language, gender, date of birth).
- Personality data – data that characterizes the learner as personality (individual features, learning style, concentration skills, personality type, collective work skills, emotional situation, attitudes).

- Pedagogical data – data which characterizes anything that a learner must learn (programs, themes, course sequence, plan).
- Preference data – data that adapts working environment for learning (language, presentation format, sound value, video speed, web design personalization).
- System experience – data that characterizes learner's earlier gained experience with computers and software used in the learning process (obtained certificates, skills in e-Learning system utilization).
- Cognitive data – data that represents reference types of the learner.
- History data – data about all learner's activities.
- Device data – data that characterizes working environment of the system user (hardware, download speed, screen resolution); learner's location, time; and devices used.
- Student knowledge at the current moment of time – data that describes student knowledge gained in the learning process.

Future work would be some kind of practical realization of the learner model using obtained results about data included in the learner model and their types.

References

[1] Behaz, A. and Djoudi, M. Adaption of learning resources based on the MBTI theory psychological types. *IJCSI International Journal of Computer Science Issues*, 9, Issue 1, No 2, (2012).

[2] Brjanik, G. and Tasso, C. A Shell for Developing Non-monotic User Modeling Systems. *International Journal of Human-Computer Studies 40* (1994), 31-62.

[3] Brusilovsky, P. Methods and techniques of adaptive hypermedia. *Spec. Issue On Adaptive Hypertext and Hypermedia, User Modeling and User Adapted Interaction*, 6, 2-3 (1996), 87-129.

[4] Brusilovsky, P. The construction and application of student models in intelligent tutoring systems. *Journal of Computer and System Scences International*, 32, 10 (1994), 70-89.

[5] Challis, M. *Portfolio-based learning and assessment in medical euducation.* AMEE Medical Education Guide No.11. 1999.

[6] Colan, O., Dagger, D., and Wade, V. Towards a Standards-based Approach to e-Learning Personalization using Reusable Learning Objects. In *World Conference on E-Learning in Corp., Govt., Health & Higher Ed.* AACE, (2002), 210-217.

[7] Colan, O., Wade, V., Bruen, C., and Gargan, M. Multi-Model, Metadata Driven Approach to Adaptive Hypermedia Services for Personalized eLearning. In *Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (2002), 100-111.

[8] De Bra, P. and Ruiter, J. P. AHA! Adaptive hypermedia for all. In *WebNet'2011, World Conference of the WWW and Internet*, AACE, (2001) 262-268.

[9] Esichaikul, V., Lamnoi, S., and Bechter, C. Student Modelling in Adaptive E-Learning Systems. *Knowledge Management & E-Learning: An International Journal*, 3 (2011), 342-355.

[10] Frias-Martinez, E., Magoulas, G., Chen, S., and Macredie, R. Automated User Modeling for Personalized Digital Libraries. *International Journal of Information Management*, 26, 3 (2006), 234-248.

[11] Garcia-Barrios, V. Personalisation in Adaptive E-Learning Systems. A Service-Oriented Solution Approach for Multi-Purpose User Modelling Systems. Dissertation in fulfilment of the requirements for academic degree, http://www.iicm.tu-graz.ac.at/iicm_thesis/vgarcia.pdf, (2007).

[12] Garlatti, S., Iksal, S., and Kervella, P. Adaptive On-Line Information System by means of a Task Model and Spatial Views. In *2nd Workshop on Adaptive Systems and User Modeling on the WWW (*1999), 55-69.

[13] Gomes, P., Antunes, B., Rodrigues, L., Santos, A., Barbeira, J., and Carvalho, R. Using Ontologies for eLearning Personalization. In *3rd Learning Conference*, Portugal (2006).

[14] Hauger, D. and Köck, M. State of the Art of Adaptivity in E-Learning Platforms. In *15th Workshop on Adaptivity and User Modeling in Interactive Systems* Halle/Saale, Germany (2007), 355-360.

[15] Hewagamage, K. P. and Lekamarachchi, R. S. Learning Patterns: Towards the Personalization of ELearning. In *5th International Information Technology Conference* (2003).

[16] Kavcic, A. Fuzzy user modeling for adaptation in educational hypermedia. *IEEE Transactions on Systems, Man, And Cybernetics*, 34, 4 (2004).

[17] Kay, J. UM: A Toolkit for User Modelling. In *Second International Workshop on User Modeling* (1990), 1-11.

[18] Kay, J. *A Scrutable User Modelling Shell for User-Adapted Interaction*. Sydney, 1999.

[19] Kay, J. The um Toolkit for Reusable, Long Term User Models. *User Modeling and User Adapted Interacton: The Journal of Personalization Reseach 4* (1995), 149-196.

[20] Kay, J. User modeling for adaptation. In Stephanidis, C., ed., *User Interfaces for All: Concepts, Methods, and Tools*. Florence, 2000, 271-294.

[21] Khirbi, M., Jemni, M., and Nasraoui, O. Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. In *Educational Technology & Society* (2009), 30-42.

[22] Kobsa, A. Generic User Modeling Systems. In *User modeling and user-adapted interaction*. (2001), 49-63.

[23] Koch, N. *Software Engineering for Adaptive Hypermedia Systems*. Verlag Uni-Druck, Munich, (2001).

[24] Kules, B. User Modeling for Adaptive and Adaptable Software Systems. In *ACM Conference on Universal Usability*,Arlington (2000).

[25] Liu, H., Salem, B., and Rauterberg, M. A survey on user profile modeling for personalized service delivery systems. In *IADIS International Conference on Interfaces and Human Computer Interaction* (2009), 45-51.

[26] Li, Q., Zhong, S., Wang, P., Guo, X., and Quan, X. Learner Model in Adaptive Learning System. *Journal of Information & Computational Science*, 7, 5 (2010), 1137-1145.

[27] Martins, A., Faria, L., Vaz de Carvalho, C., and Carrapatoso, E. User Modeling in Adaptive Hypermedia Educational Systems. In *Educational Technology & Society*, 11, 1 (2008), 194-207.

[28] Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I., and Wade, V. Adaptive educational hypermedia systems in technology enhanced learning: a literature review. In *2010 ACM Conference in Information Technology Education*, Midland, USA (2010).

[29] Nebel, I., Smith, B., and Paschke, R. A user profiling component with the aid of user ontologies. In *Workshop Learning - Teaching - Knowledge - Adaptivity (LLWA 03),* Karlsruhe, (2003).

[30] Nguyen, L. and Do, P. Learner Model in Adaptive Learning. In *World Academy of Science, Engineering and Technology* (2008), 396-401.

[31] Orwant, J. *Heterogenous learning in the Doppelganger user modeling system*, (1995), 107-130.

[32] Paiva, A. and Self, J. TAGUS: A User and Learner Modeling System. In *4th International Conference on User Modeling*, Hyannis (1994), 43-49.

[33] Paramythis, A. and Loild-Reisinger, S. Adaptive Learning Enviroments and eLearning Standarts. *Electronic Journal on e-Learning*, 2, 1 (2004), 181-194.

[34] Queiros, R., Olivera, L., Leal, J., and Moreira, F. Integration of ePortfolios in Learning Management Systems. In *International Conference Computational Science and Its Applications - ICCSA 2011,* Santander, (2011), 500-510.

[35] Riad, A., El-Minir, H., and El-Ghareeb, H. Review of e-Learning Systems Convergance from Traditional Systems to Services based Adaptive and Intelligent Systems. In *JCIT* (2009), 108-131.

[36] Rich, E. Stereotypes and User Modeling. *User Models in Dialog Systems* (1989), 35-51.

[37] Rich, E. User Modeling via Stereotypes. *Cognitive Science*, 3, 4 (1979), 329-354.

[38] Romero, C. and Ventura, S. Educational data mining: A review of the state-ofthe-art. *IEEE Transactions on Systems, Man, And Cybernetics, part C: Applications and Reviews*, 40, 6 (2010), 610-618.

[39] Shute, V. and Towle, B. Adaptive e-learning. *Educational Psyhologist*, 38, 2 (2003), 105-114.

[40] Somyürek, S. Student Modeling: Recognizing the Individual Needsof Users in e-Learnig Enviroments. *International Journal of Human Sciences*, 6, 2 (2009), 429-450.

[41] Sosnovsky, S. and Dicheva, D. Ontological technologies for user modeling. *Iternational Journal Metadata, Semantics and Ontologies*, 5, 1 (2010), 32-71.

[42] Van Wasel, M. and Prop, A. The infulance of portfolio media on student perceptions and learning outcomes. In *Student Moblity and ICT: Can E-LEARNING overcome berriers of Life-Long learning?,* Maastricht, 2008.

[43] Vasilyeva, E., Pechenizkiy, M., and Puuronen, S. Knowledge management challenges in web-based adaptive e-learning systems. In *5th International Conference on Knowledge Management,* Springer, (2005), 112-119.

[44] Vergana, H. PROTUM: A Prolog Based Tool for User Modeling. *WG Knowledge-Based Information Systems, Department of Information Science* (1994).

Part 2. Abstracts

Abstracts of Papers in Post-Proceedings

1. Michael Soffner, Norbert Siegmund, Marko Rosenmüller, Janet Feigenspan, Thomas Leich and Gunter Saake. A VARIABILITY MODEL FOR QUERY OPTIMIZERS

By adopting to more domains, database management systems (DBMSs) increase their functionality continously. This leads to DBMSs that include often unnecessary functionality, which decreases performance. A result of this trend is that new specialized systems arise that focus only on a certain application scenario but often reimplement already existing functionality. To overcome overbloated DBMSs, we propose to introduce variability in DBMS implementations that allows users to select only needed functionality for a specic application scenario. In this paper, we focus on the query optimizer as it is a key component of DBMSs. We describe the potentials of tailoring query optimizers. Furthermore, we analyze common and diering functionality of three query optimizers of industrial DBMSs (SQLite, Oracle, and PostgreSQL) to create a variability model for query optimizers that can be used as a basis for future variability-aware implementations.

2. Igor Epimakhov, Abdelkader Hameurlain and Franck Morvan. GEOLOC: ROBUST RESOURCE ALLOCATION METHOD FOR QUERY OPTIMIZATION IN DATA GRID SYSTEMS

Resource allocation (RA) is one of the key stages of distributed query processing in the Data Grid environment. In the last decade were published a number of works in the field that deals with different aspects of the problem. We believe that in those studies was given insufficient attention to such important aspects as allocation space definition and criterion of parallelism degree determination. In this paper we propose our method of RA that extends existing solutions in those two points of interest and resolves the problem in the specific conditions of the large scale heterogeneous environment of Data Grid. Firstly, we propose to use a geographical proximity of nodes to data sources to define the AS. Secondly, we present the principle of execution time parity between read and join operations for determination of the parallelism degree and generation of load balanced query execution plan. We conducted an experiment that proved the superiority of our GeoLoc method in terms of response time over the RA method that we chose for the comparison. The present study provides also a brief description of existing methods and their qualitative comparison with the proposed method.

3. Boris Novikov, Elena Mikhaylova, Ekaterina Ivannikova and Alice Pigul. MINING LOGS FOR LONG-TERM PATTERNS

In this work we made an approach for data storage system optimization. Most high-capacity storage systems consist of several devices. These devices may have different performance. The goal is to control data placement in such way that data are moved to faster devices just before they are expected to be intensively used. To accomplish we would like to find long term data access patterns. However the high level application logic and schedules are not available at the storage system level. Our approach is to use log mining to identify data access patterns. If the system has information about data that will soon be required for processing, it is possible to prepare the data by transferring them to a faster storage parts. We analyze the database log files containing the history query executions and identify repeating query groups. Our hypothesis is that this query groups are closely related with meaningful business processes of the application. These groups are very likely related to the business process. Knowing the business processes, we can determine the data they need.

In this paper we offer the algorithm for query groups' detection that and describe the parameters affecting algorithm efficiency. Also we describe the algorithm for periods identifying for detected query groups.

Testing the algorithm on real production data showed that the proposed algorithm identifies more than 60% of known business processes.

4. Benameur Ziani and Youcef Ouinten. COMBINING DATA MINING TECHNIQUE AND QUERY FREQUENCIES FOR AUTOMATIC SELECTION OF INDEXES IN DATA WAREHOUSES

Index selection is an important part of physical database design. Its goal is to select an appropriate set of indexes to minimize the cost for a given workload under storage constraint. However, selecting a suitable configuration of indexes is a difficult problem to solve. The problem becomes more complex for indexes defined on multiple tables such as bitmap join indexes, since it requires the exploration of a much more search space. Studies dealing with the bitmap join indexes selection problem mainly focused on proposing pruning solutions of the search space by the means of data mining techniques or heuristic approaches. So far, the data mining based approaches have used closed frequent itemsets to reduce the search space for the selection process. These approaches have two notable shortcomings. Firstly, they generate a huge number of indexes with a lot of redundancy that it is very difficult to manage according to the system limitation (number of Indexes per table, storage space constraint). Secondly, when they construct the extraction context for mining frequent sets of attributes, they have used indexable attributes only once for each query in the workload which does not reflect the importance of a given query in the workload. Indeed, the queries in a workload are unlikely to have the same probability of being requested. To overcome these imitations, we propose to combine maximal frequent itemsets and query frequencies to improve the quality of generated indexes. This paper describes an approach that refines the index selection process, incorporating query frequencies in the extraction context for mining frequent set of attributes. We experimentally prove that our approach reduces the storage space and improves the quality of the recommended indexes.

5. Janari Põld, Tarmo Robal and Ahto Kalja. ON PROVING THE CONCEPT OF AN ONTOLOGY AIDED SOFTWARE REFACTORING TOOL

Through years more and more software is produced. The quality of software architecture however has an important role in systems exploitation, as it determines the maintainability and extensibility of an application. Recently more emphasis is put on quality of the design, so that new features can be added with ease. To preserve code readability and extensibility, software architecture must be refactored from time to time to cope with the modifications. Nevertheless, reviewing the whole source code is time consuming and does not return any surplus, thus it is often skipped, causing the software architecture to decay in time over several modifications and making it harder to add new functionality in the future. An automated method of recognizing "bad" code would help to solve some of the issues. In this article the authors propose a concept of a refactoring tool, which uses ontology to find “smelly” design and tackle the aforementioned problems. Several aspects of the tool are discussed – how it works and how it can be used to improve the software architecture, thus augment the quality.

6. Kārlis Čerāns, Renārs Liepiņš, Jūlija Ovčiņnikova and Arturs Sprogis. ADVANCED OWL 2.0 ONTOLOGY VISUALIZATION IN OWLGRED

Intuitive ontology visualization is a key for their learning, exchange, as well as their usage in conceptual modeling and semantic database schema design. OWLGrEd is a visual tool for compact graphical UML-style rendering and editing of OWL 2.0 ontologies. We describe here the extensibility features for OWLGrEd that allow tailoring the editor for specific ontology-based modeling needs, including custom entity annotation visualizations and description of integrity constraints for semantic database schemas. We discuss the application of concrete OWLGrEd extensions in the context of ontology-centered information system engineering.

7. Uldis Donins. FORMAL ANALYSIS OF PROBLEM DOMAIN WORKFLOWS

The formal foundation of Topological functioning model (TFM) makes it as a powerful tool to analyze the functioning of a problem domain and to formally relate problem domain artifacts with the artifacts that should exist in solution domain. TFM captures system functioning specification in the form of topological space consisting of functional features and cause-and-effect relations among them and is represented in a form of directed graph. The functional features together with topological relationships contain the necessary information to create diagrams of other type, e.g., Activity or Class diagrams. To specify the behavior of system execution a new artifact is added to the TFM – the logical relations. The presence of logical relations within TFM denotes forking, branching, decision making, and joining during the functioning of the system. Thus it is needed to identify and carefully analyze logical relations within TFM in order to have all the necessary information to transform it to diagrams of other type. This paper gives the formal method of transforming TFM into Activity diagram together with an example of such transformation.

8. Janis Barzdins, Edgars Rencis and Agris Sostaks. TOWARDS HUMAN-EXECUTABLE BUSINESS PROCESS MODELING

There are many organizations, whose everyday life involves lots of tasks performed, or let us say executed, by lots of different people. Since nowadays processes have become much more complex, a big challenge for humans is to even understand what, when and how have to be done in order to reach their goals. Business process models are frequently used in organizations to make the process understandable to performers and to alleviate their work by connecting the process to organization's information system thus making processes human-executable. However, while developing a solution, there are usually only two extremes to choose from – either we use an all-in-one solution for describing process steps or we develop a domain-specific process modeling language from scratch. In this paper we propose the golden mean – a good base for domain-specific process modeling languages and appropriate tooling to be used in a big portion of related organizations and relatively easily integrated into their information systems. We define, what is meant to be “good” by binding the process language base with the natural language generator. We also demonstrate the approach on a case study of a process modeling language for the University of Latvia.

9. Dejan Lavbič, Slavko Žitnik, Lovro Šubelj, Aleš Kumer, Aljaž Zrnec and Marko Bajec. TRAVERSAL AND RELATIONS DISCOVERY AMONG BUSINESS ENTITIES AND PEOPLE USING SEMANTIC WEB TECHNOLOGIES AND TRUST MANAGEMENT

There are several data silos containing information about business entities and people but are not semantically connected. If in integration process of data sources trust management is also employed than we can expect much higher success rate in relations discovery among entities. Majority of current mash-up approaches that deal with integration of information from several data sources omit or don't fully address the aspect of trust. In this paper we discuss semantic integration of personal and business information from various data sources coupled with trust layer. The resulting system has higher and more defined solidity while trust for single entity and also for data source is defined. The case study presented in the paper focuses on integration of personal information from data sources mainly maintened by government authorities which have higher trustability than information from social networks, but we also include other less trusted sources. The developed SocioLeaks system allows users traversal and further relation discovery in a graph based manner.

10. Erika Asnina, Janis Osis and Asnate Jansone. FORMAL SPECIFICATIONS OF TOPOLOGICAL RELATIONS

The paper discusses application of the topological functioning model (TFM) of the system for its automated transformation to behavioral specifications such as UML Activity Diagram, BPMN diagrams, scenarios, etc. The paper addresses a lack of formal specification of causal relations between functional features of the TFM by using inference means suggested by classical logic. The result is reduced human participation in the transformation as well as additional check of analysis and specification of the system.

11. Elena Sivogolovko. THE INFLUENCE OF DATA QUALITY ON CLUSTERING OUTCOMES

Relationship between Clustering and Data Quality has not been thoroughly established. It is usually assumed that input dataset does not contain any errors or contains some "noise", and this concept of "noise" is not related to any Data Quality concept. In this paper we focus on the four most commonly used data quality dimensions, namely accuracy, completeness, consistency and timeliness. We evaluate the impact of data quality on clustering outcomes using denitions and constructs of these quality dimensions. Four dierent clustering algorithms and ve real datasets were selected to show the interaction between data quality and cluster validity.

12. Vitaly Zabiniako. VISUALIZATION OF GRAPH STRUCTURES WITH MAGNETIC-SPRING MODEL AND COLOR-CODED INTERACTION

In this paper author provides description of the original approach for visual analysis of data represented with general graphs, based on modification of magnetic-spring model and color-coded cognitive manipulation with graph elements. The theoretical background of magnetic fields in application to graph drawing is presented along with discussion of appropriate visualization techniques for improved information analysis and comprehension. Usage of other existing graph layout strategies (e.g. hierarchical, circular) in conjunction with magnetic-spring approach are also considered for improved data representation capabilities. A concept of integrated virtual workshop for graph visualization is introduced which relies on aforementioned model and can be used in GVS (Graph Visualization Systems). A case study of application of proposed approach is presented along with conclusions of its usability and potential future work in this field.

13. Janis Grundspenkis and Antons Mislevics. MOBILE AGENTS FOR INTEGRATING CLOUD-BASED BUSINESS PROCESSES WITH ON-PREMISES SYSTEMS AND DEVICES

Business Process Management Systems (BPM systems) are used to control, analyze and manage business processes in organizations. BPM systems help to reduce the amount of administrative effort and focus on the processes which add value. Nowadays, moving towards cloud-based Software-as-a-Service (SaaS) architecture, some additional requirements for successful BPM implementation are identified. One of the main challenges is how to ingrate SaaS BPM systems with existing on-premises systems, data sources and devices. In this paper, mobile agents are proposed as the technology addressing this new challenge. A mobile agent is a composition of computer software and data which is able to migrate from one device to another autonomously and continue its execution on the destination device. The paper starts with and overview of SaaS BPM and existing approaches to address SaaS integration challenges. Then, the concept of mobile agents is described, and the idea of how mobile agents may be used in SaaS BPM integration scenarios is presented. The paper is continued with a comparison of widely used integration approaches with proposed mobile agents based mechanism. Finally, a newly proposed architecture is presented in a prototype, outlining its advantages and proposing directions for future research.

14. Tarmo Robal and Ahto Kalja. APPLYING USER DOMAIN MODEL TO IMPROVE WEB RECOMMENDATIONS

The enormous amount of information available over the Internet has forced users to face information overload while browsing the World Wide Web. Alongside with search engines, recommender systems and web personalization are seen as a remedy to this problem, since users are browsing the web according to their informational expectations while having a sort of implicit conceptual model in their mind. The latter is partially shared with other site visitors. In this paper we apply ontological modeling of anonymous ad-hoc web users' behavior to improve online user action prediction for web personalization via recommendations.

15. Riina Maigre, Pavel Grigorenko, Hele-Mai Haav and Ahto Kalja. A SEMANTIC METHOD OF AUTOMATIC COMPOSITION OF E-GOVERNMENT SERVICES

It is hard to automatically find a semantically meaningful web service composition over a huge collection of web services available on the web. However, recent results in semantic web service research and technology could be effectively used within some specific domains. E-government is one of the sectors that need horizontal integration. Therefore, semantic web services and their composition become necessary and applicable in this domain. The paper proposes a semantic method of automatic composition of e-government services. It uses domain ontologies presented in OWL, semantic web services described in SAWSDL, quality of service (QoS) characteristics, ontology reasoning and AI planner in order to automatically provide service plans that could be presented in BPEL for execution. The approach is motivated by a case study from the domain of the Estonian state information systems.

16. Kristiina Kindel, Urve Venesaar and Merli Reidolf. COMMUNICATION CHANNEL CHOICE BETWEEN ENTERPRISES AND GOVERNMENT

Communication channel choice is the use by enterprises of one media channel compared to another (Reddick & Turner, 2012). Channel choice has been studied in media in the use and gratification literature (Kaye & Johnson), and the question whether old media are driven out of existence by new media or the importance of choosing right media for communication has been a concern in academic and industrial research (Nguyen &Western, 2006; Lengel & Daft, 1989; Vassilakis, Lepouras Halatsis, 2007). Despite of fast increase in the use of e-government services there still exists a need of enterprises to contact with government via traditional channels. The literature on why enterprises initiate contact with government through different communication channels has not got much attention.

The aim of current article is to identify the factors influencing the enterprises' choice of communication channels with government comparing e-government to traditional service delivery channels such as the phone, mail, fax or visiting a government office. The study examines factors that explain the choice of channels according to the reasons for communication with government as well as depend on the characteristics of enterprises (e.g. sector, size, ownership, location, strategic choices). When focusing on the online portals of government institutions the impact of external factors influencing the use of e-government services will be analysed. In addition, the enterprises opinion about their satisfactory experience with public service delivery and benefits as well as problems connected with the use (or not use) of e-government services will be used to determine their impact to the choice of communication channels.

The main research questions are: 1) What factors explain enterprises' choice of communication channels with government; 2) What factors could impact the increase of the use of e-government services.

The article, through logistic regression of enterprises' opinion survey in Estonia and Germany is assessing the most commonly used communication channels depend on the nature of enterprises' interaction between government and other characteristics of enterprises, and their experience with using e-government services. The results of analysis should show the reasons for using multiple channels for conducting with government, and whether there will be possibilities for increasing the use of e-government services in enterprises.

17. Evari Koppel and Raimundas Matulevicius. An Evaluation Framework for Software Test Management Tools

Software testing has proven its value for software development increasingly over the last decade. With the recognition of the benefits of software testing, several software test management tools (TMT) have emerged on the market. Although there exist different approaches, there is no method for a systematic TMT assessment. This is a problem because to our knowledge, evaluating TMT is rather the subjective task, heavily depending on the evaluators' opinions rather than based on the objective approach. The same problem applies when test managers are asked to evaluate whether their currently used TMT meets the company's expectations. In this paper based on the survey performed among Estonian testing practitioners, we deliver a TMT evaluation framework. The paper applies structured approach by performing a literature study on software testing processes, existing TMT market research, and mapping together the identified test activities and test artifacts. The results help formulate and design the online questionnaire and perform a TMT survey in the Estonian IT companies. Based on this survey results, a framework for evaluating TMT software is created. Such a framework could potentially help companies to measure the TMT suitability to company's goals and to decrease subjectivity of the TMT assessment. The framework also provides test and project managers the understanding whether their current TMTs meet the company's expectations.

18. Edgars Diebelis and Janis Bicevskis. SOFTWARE SELF-TESTING

The Paper presents an overview of the results of 5 years of research in the field of self-testing. In 2007, self-testing was defined as one direction of smart technologies, a common idea of which is the desire to fit software with features of living beings: abilities to adapt to changing external environment, to optimise themselves and to defend themselves against threats. The purpose of self-testing is to provide a possibility to verify that the software is working correctly at any point of its life cycle. The research was carried out in several stages: at first, the concept and functionality of self-testing and its applicability in various software operating environments were defined; it was followed by implementing the self-testing functionality by integrating testing support options into the software developed. After that, the self-testing concept was compared against the possibilities offered by traditional testing support tools and implemented in an actual banking information system, and the efficiency of self-testing options was evaluated. The final conclusions drawn are: self-testing offers a number of advantages in achieving the software quality at comparatively low costs, at the same time ensuring the same functionality as provided by conventional testing support tools.

19. Guntis Arnicans, Dainis Romans and Uldis Straujums. SEMI-AUTOMATIC GENERATION OF A SOFTWARE TESTING LIGHTWEIGHT ONTOLOGY FROM A GLOSSARY BASED ON THE ONTO6 METHODOLOGY

We propose a methodology of semi-automatic obtaining of a lightweight ontology for software testing domain based on the glossary “Standard glossary of terms used in Software Testing” created by ISTQB. From the same glossary many ontologies might be developed depending on the strategy for extracting concepts, categorizing them, and determining hierarchical and some other relationships. Initially we use the ONTO6 methodology that allows identification of the most important aspects in the given domain. These identified aspects serve as the most general concepts in taxonomy (roots in the concept hierarchy). By applying natural language processing techniques and analyzing the discovered relations between concepts, an intermediate representation of lightweight ontology is created. Afterwards the lightweight ontology is exported to OWL format, stored in the ontology editor Protégé, and analyzed and refined by OWLGrEd – an UML style graphical editor for OWL that interoperates with Protégé. The obtained lightweight ontology might be useful for building up a heavyweight software testing ontology.

20. Stanislovas Norgėla, Julius Andrikonis and Arūnas Stočkus. QUALITATIVE REASONING ABOUT SPACE WITH HYBRID LOGIC

This article describes the way to employ hybrid logic H(@,↓) in the analysis of qualitative spatial information. Moreover, it shows how the complexity of model checking algorithm is derived using the Kripke structure of qualitative spatial information and the query, which is presented as a formula of hybrid logic.

21. Laura Savičienė. Modeling Operationalization of Normative Rules in Decision Support for Aircraft Approach/Departure

This research is focused on norm operationalization in aeronautics domain. The investigated paradigm can be described as: from legal norms to technical rules in the artifact. Normative requirements (norms) for the aircraft trajectories are extracted from the flight rules and airport procedures. These norms are operationalized in a decision support system (DSS). An example of a normative rule: keep 3 degree descent angle while landing and hold restrictions of altitude and geography which is depicted in the approach chart. The decision support is based on evaluation of risk to violate the normative requirement. The following risks are modeled: trajectories' conformance with the flight rules, safe distance between aircraft, wake vortex separation and avoidance of dangerous substances in the atmosphere. The DSS is for the air traffic controller (not pilot) and must respond in real time. A DSS system provides surveillance, evaluates and recommends, whereas the human controller takes a decision.

22. Juris Ivanovs and Kriss Rauhvargers. Handling server-side software versioning: the "smart technology" approach

Deploying new versions of server-side software is similar to deploying new versions of desktop software, however it is considered more complex and time consuming. Therefore, if new versions are released frequently and they need to be deployed to many servers, doing the work manually may lead to several problems - errors due to incorrect deployments, misconfigurations and considerable amount of time spent on routine tasks. This paper is a study of methods used for desktop software versioning in order to apply them to server-side software needs. The main focus was set on server-side software that is based on PHP and Oracle technology, however solutions where sought that could be used for other serverside technologies as well, e.g., ASP.NET, Java and Ruby. As a result, a solution was created and applied in a real-world scenario that helps handling server-side software versioning by automating builds of new versions, deployment and validation processes.

23. Rudolfs Bundulis and Guntis Arnicans. Architectural and Technological Issues in the Field of Building High-Resolution Display Walls

Currently there is a rising need to lay out a vastly growing amount of information and supersize working areas for collaboration and presentation needs. The hardware side is not able to catch up with the needs – display surfaces are still limited either in size or resolution and are not capable to offer a homogenous large scaled display with a high resolution to present the needed amount of information. This issue is tackled by constructing a multiple display wall that has a tiled display surface where the resolution is high enough since it sums up the individual resolutions of each tile. But as this solution is also limited to the number of video cards in the computer and their ability to feed multiple display targets and different, there are ongoing studies to understand how to cope with the current limitations on bandwidths by altering the architecture of the solution. This paper summarizes the current limitations and cost-effectiveness of display wall environments and proposes ideas for alternate solutions.

24. Arturs Sprogis and Janis Barzdins. SPECIFICATION, CONFIGURATION AND IMPLEMENTATION OF DSL TOOL

A new specification method for DSL and DSL tools is proposed. The method is based on an advanced stereotype mechanism. A special feature of the proposed method is a precise definition of the extension mechanism for realization of non-standard features of DSL tools. In conclusions the architecture of a DSL tool building framework based on the proposed specification method is described.

25. Inga Zilinskiene and Saulius Preidys. A MODEL FOR PERSONALIZED SELECTION OF A LEARNING SCENARIO DEPENDING ON LEARNING STYLES

This paper deals with one of Technology Enhanced Learning (TEL) problems - the personalized selection of a learning scenario. Personalization is treated here as appropriateness of a learning scenario to preferences of a particular student, mainly, his/her learning style. This paper proposes an extended approach to modelling learning scenario selection based on preferences of a student‘s learning style. An ant colony optimization algorithm is modified and applied. In order to give a theoretical background the main conceptions of personalization, learning scenario and learning style are briefly presented. The aim of this paper is twofold. First, data mining technique to obtain a student‘s learning style is presented, second, a model for personalized selection of a learning scenario is proposed.

26. Oskars Rasnacs and Maris Vitins. AN INFORMATION SYSTEM TO LEARN CHARACTERISTIC SETS OF WORDS AND TO EXAMINE KNOWLEDGE IN STATISTICS

The authors have found that many students in the fields of health care and the social sciences, as well as practicing specialists, have problems when writing bachelor’s or master’s theses or other scholarly publications when it comes to taking decisions on the most appropriate data processing methods in their work. The authors have studied and analysed the theses and papers that have been produced, as well as the data processing methods that are indicated therein. Aspects of statistics are discussed in various areas of specialisation and in various courses. This means that students usually obtain a lot of information that is useful, but very hard to remember; they do not learn about schemes related to how the information can be brought to bear. This paper is based on the question of what students and practicing specialists must remember if they hope to find the necessary information from various sources (the Internet, the literature) to make independent decisions about the acceptance of appropriate data processing methods and about the implementation of those methods. The authors have found that there are many situations in the area of data processing which can be classified in different ways, and course instructors have divergent views about the most appropriate method for each situation. At the same time, each situation is in line with several sets of characteristic words. Because software package management teams, assistance systems and educational literature are all usually in English, it is recommended that students learn the terminology in English irrespective of the language of the course which they are taking. The authors led a working group to design an information system in which each course instructor can implement a classification of data processing methods which is acceptable to him or her, also coming up with characteristic sets of words which are in line with the situation, as well as appropriate examples of data files. There is no denying that it would be more useful for students to work with data from real patients, but legal acts make that impossible. That is why the authors have addressed the issue of generating data on the basis of statistical indicators from scholarly publications. Students and specialists can use this information system in the educational and the test regime. In the education regime, the generated data files and corresponding sets of characteristic words can be examined. The test regime examines knowledge about the sets of characteristic words. The proposed information system has been tested in a traditional educational process at the university level, as well as in individual training sessions. Participants in the tests were tested and surveyed via a questionnaire. The results proved the effectiveness of the approach and the system.

27. Svetlana Kubilinskienė and Valentina Dagienė. METHODOLOGICAL DIGITAL RESOURCES: HOW WE CAN HELP EDUCATORS TO FIND THEM MORE EFFECTIVELY

The paper deals with digital resources in education and mainly focus on an approbation of the extended metadata model of digital learning resources. The model has been developed by covering methodological resources and learning method objects in order to increase their accessibility and usage in teaching process. The key purpose of methodological resources is to render conditions for teachers to share the professional experience, to spread methodological novelties, to help students and their parents to join the training and learning process more effectively. Different ways of choosing and combining learning methods obligate teachers first of all to know and estimate them, in line with the requirements posed to the contemporary school. Effective learning resource search and browsing possibilities can be realized only if standardized metadata are used. The metadata are the essential part of information infrastructure which is necessary while establishing order in internet chaos by using descriptions, classifications and structures which are helpful in creating more power and useful information repositories. At the moment the extended metadata model is implemented in the Lithuanian learning object metadata repository prototype. The paper focuses mainly on the results of an experimental approbation of the metadata model.

Part 3. Materials of Doctoral Consortium

Trends of the Usage of Adaptive Learning in Intelligent Tutoring Systems

Jānis DĀBOLIŅŠ [1]
Faculty of Computer Science and Information Technology, Institute of Applied Computer Science, Department of Systems Theory and Design, Riga Technical University, Phone: +371 67089529

Abstract. In this paper general tendencies of adaptive learning are described, which are becoming more and more common in the time when IT technologies and use of internet as well as development of web systems is live topic in researches about technology enchanted learning with the aid of intellectual learning systems. Adaptive learning analysis, intellectual tutoring system description and general ITS structure and general development principles are described based on the bibliography. ITS modules, ITS collaboration with learner, necessity of feedback and existing adaptive learning methods have been described.

Keywords. Intelligent tutoring systems, interactive learning environments, adaptive learning, agent

Introduction

Computer usage in learning is connected with IT development in general; it began at the end of 1950's when machines, which are considered primitive nowadays, were constructed for programmed training. Currently such technologies are used very widely, for example, e-format information literature, virtual training systems and environments, self-appraisal tests, technology enchanted learning as well as animated tutorials.

Improvement of IT technologies, expansion of internet and popularization of web technologies have enabled technology enhanced learning introduction in adoption of general matters and acquaintance of specialized problems. Thus researches on adoption of learning contents into e-environment have become more necessary as a result of such development [21, 9, 1]. Besides technology enhanced learning lets one to pick the place and time where and when to study, which is great advantage compared to traditional full-time education [16].

Identical educational tools and learning methods may not be effective or is less effective for all students. It is possible to make the learning materials more flexible, modify educational approach according to competence and temper of the student and learning tasks, using Intelligent Tutoring Systems (ITS). It is possible to achieve better learning results when fitted educational aids are used for individual needs of every student [13]. Solutions are hunted in diversiform approaches, where ITS already takes important place, as to a certain extent provides flexibility, adaptation, etc. Flexible approach in learning is achieved by usage of several learning strategies, which are

[1] janis.dabolins@rtu.lv

implemented according to student's progress in learning of the material. Such flexible learning system for industrial purposes has already been described in 1998 [4], where adaptive learning system is described with several learning strategies from several agents.

Further researches [21, 13, 1] are focused on development of adaptive learning system [11]. General adaptive learning tendencies are analysed in this paper and insight onto further researches are given.

1. Analysis of Adaptive Learning

As volume of learning materials is great in e-environment, it is hard for consumers and information seekers choose the right materials, thus it would be necessary to create adaptive learning systems. Adaptive learning may be defined as ‘‘the process of generating a unique learning experience for each learner based on the learner's personality, interests and performance in order to achieve goals such as learner academic improvement, learner satisfaction, effective learning process and so forth" [21].

Learning, methods and student's reaction to learning is not easily definable and describable. Learning methods and successful learning results may not be defined as particular thing, order of matters, sequence of events and thus guarantee successful outcome (student is trained and knows everything that he/she needs). That is the reason why adjustment of learning content to the student's competence is considered to be an open system problem, where adaptation capacity of the student, collaboration with learning environment as well as preferable result of system action plays great importance (it is even hard to define the necessary outcomes, as it is not possible fully evaluate persons' gained knowledge and its system) [19]. In the same way it is hard to define where system's effect on person comes to an end.

When analysing person's behaviour changes as a result of experience, comparing student's behaviour in time period 1 to behaviour in time period 2 (Figure 1) – if behaviour is different in the same circumstances, we may consider that learning has taken place. Such analysis of behaviour would give information on necessity of further training for integrated learning agent.

In the development of learning systems it is necessary to take into account both persons' needs and requirements, as well as resources of information technologies. It is also necessary to evaluate the use of didactic materials – cover learning theory matters, develop training content plan, training methods and organizational forms, all these operations would be arranged and subordinated according to their possible

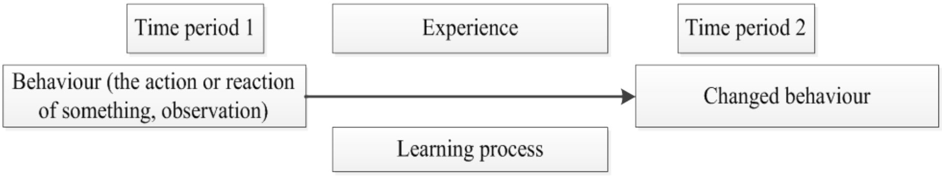


Figure 1. Reproduction of learning [5]

formalization forms for depicting and use in ITS. It is required to include student's reaction to the learning material and analyse the chances of learning outcome stimulation.

In natural environment teacher may pick up each student's individual abilities, character, reaction to learnable material and depending on these parameters, form more or less individualized task set for each individual. It is necessary to integrate analyse mechanisms and reactions to imitate or overcome natural environment achievements. On the other hand if the teacher is located in front of large student group, he/she is not able to catch each student's abilities and necessities, thus e-learning advantage is silhouetted in this case. In e-learning system it is easier to analyse each individual needs [1]. When adjusting training to individual needs, following aspects should be included: individual intellectual abilities, perceptive type, preliminary knowledge, desire for learning, motivation for achievements, self-suggestion [1]. Currently researches are based on development of adaptive systems according to two approaches: 1. adaptivity – system structures didactic material combination according to knowledge about the student; 2. adaptability – system adjusts learning content responding to change in student's knowledge during the use of system [21]. Studies are carried out on development of adaptive learning materials, based on these two approaches.

2. Intelligent Tutoring Systems

Technology enchanted learning is learning where self-motivation, communication, efficiency and technologies are used [16]. ITS is system and technologies where adaptive learning technologies are used, which help individualize and personalize learning process according to individual character and needs [8], analyse knowledge of the theme, student's mood and emotions (human decisions are based not only on analysis of various possibilities and resulting signs, but also on emotions) [18, 2], as well as learning style, typically ITS is constructed as multi-agent systems [3].

There are three main approaches on ITS development [13]:

1. The sequence of curriculum is made so that the student may easily adapt himself/herself, besides material is demonstrated to the student just when he/she needs it.
2. ITS gives detailed feedback to the learner on the imperfect or false solution, helping to learn from ones mistakes.
3. Problem solving methods - little help is provided to the learner, so the right solution is achieved.

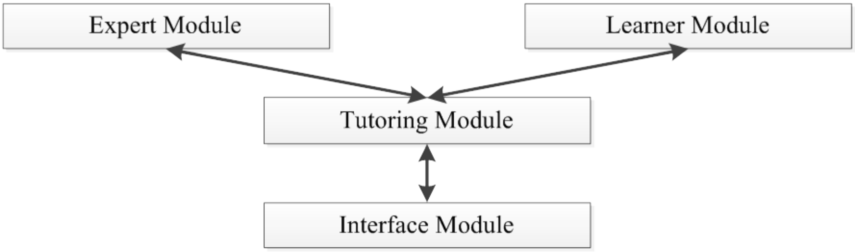


Figure 2. General ITS structure

General ITS structure is given in Figure 2. It is possible to create constructive multi-agent system for adaptive learning, when describing and formalizing information on person's knowledge accumulation and interpretation types. ITS includes both student/learner module and expert module, both of those mutually interacts with tutoring module, which is available for users in user interface. Operation of agents and adaptation to learner takes place in learning/tutoring module, where system receives information about user's activities, acquired knowledge and skills [18, 16, 14].

The necessary information for this system may be supplemented in the expert module, but the information about student is stored in the learner module. Both of these modules interactively serve as information storage for agents, thus operation of modules is defined as follows:

- Learner module is accumulative type – it stores information about student's actions, progress and test results.
- Expert module contains information about learning content. It may be supplemented and in collaboration with student module perform changes in order to bring together learning module action with solving of students' problems in possibly short period of time.
- Tutoring module cooperates with both previous mentioned modules for gaining the material as well as determines the accuracy of student's solutions, based on information contained in the expert module. Module contains algorithms as well, which performs help function for student during learning process.
- Operation in e-learning system is done in the user interface.

Agents integrated in ITS should be able to express attention, adapt themselves to abilities, learning speed, needs and necessities of the leaner [12]. These qualities are attained by integration of fuzzy logic elements into the agent, using their response in planning and expression of emotions in reaction to user's activities.

To achieve high efficiency ITS collaboration with student, it is necessary to integrate natural language processing (NLP) techniques, as a result agents integrated in the system may analyse student's answers qualitatively [15]. If feedback between ITS and learner is provided, better training results are achieved [10, 3, 17], if ITS instantly corrects the mistakes made by learner or shows the right way how to avoid the mistake, knowledge of the student becomes deeper and wider [20, 9]. The knowledge assessment agent using the comparison algorithm based on graph patterns compares a teacher's and a learner's concept maps and assigns score for submitted solution [7]. IKAS (Intelligent Knowledge Assessment System – developed at the Department of Systems Theory and Design of Riga Technical University) basic concept is exercises for knowledge assessment, where natural language processing is not necessary for computerized assessment, as well as teaching staff involvement in evaluation process of competence. Idea map exercises are different and thus are adjustable to different knowledge assessment needs (more simple and complicated exercises). Besides variety of exercises founded in idea maps and arranging possibility after level of difficulty, allows to adjust them to adaptive knowledge assessment [3]. But the system COMPASS (COncept MaP ASSessment and learning environment) is a discipline-independent concept mapping learning environment, developed at the Educational & Language Technology Laboratory of the Department of Informatics & Telecommunications at the University of Athens. The analysis of the map is based on the assessment of the propositions according to specific criteria, such as completeness, accuracy, superfluity, missing out and non-recognisability, results into the identification of specific error categories (e. g. incomplete relationship, incorrect concept, superfluous relationship, missing proposition), and is discriminated in the qualitative and quantitative analysis. The qualitative analysis is based on the qualitative characterization of the errors and aims to contribute to the qualitative diagnosis of student's knowledge; that is student's incomplete understanding/ beliefs and false beliefs. The quantitative analysis aims to evaluate student's knowledge level and is based on the weights assigned to each error category as well as to each concept and proposition that appear on expert map, reflecting their degree of importance. Pre-defined weights for error categories are supported; the teacher has the possibility to personalize the assessment process and configure the weights [6].

3. Conclusions

In former studies e-learning systems mostly have been formed as systems for information supply, where lectures or distance learning materials are available in e-format. Popularity of internet and web technologies have made e-learning development possible, though available operating systems still are "teacher's" systems where only little attention is paid to learner's needs and abilities [13]. In the latest researches target has been set to individualized e-learning system, directed to the student. Technology enhanced learning have adaptation gaps, where one of the solutions is ITS – where realization of adaptation is made, so that ITS is brought nearer to the person's ability to learn. Currently feedback role is developed the most in ITS, IKAS and COMPASS.

When analysing existing ITS we may conclude that their development is progressing, but still it is not at the level where ITS fully realizes adaptive training – technologies which let course to comply with students' knowledge level and preferences automatically. Adjustment of learning content and didactic materials to student's needs is object of many researches, but trends differs – several researches tries to develop standardized learning content, which could be modified according to progress of the student, others in their turn are directed to personalization of learning content, adaptation directly to student's goals and achievable results. All of these approaches have their drawbacks – first, most of the results and elaboration of these researches are not compatible with existing learning standards, second – ITS that adjusts to the interests of learner may omit important parts of the learning content. Irrespective ITS drawbacks, elaborating and research is very current topic because of e-learning privileges, thereby we may consider development in the direction of ITS to overcome their imperfections and drawing nearer to real setting learning or even gaining advantage over these.

Acknowledgement

This work has been supported by the European Social Fund within the project "Support for the implementation of doctoral studies at Riga Technical University".

References

[1] Y. Akbulut and Ç. S. Çardak, Adaptive educational hypermedia accommodating learning styles: a content analysis of publications from 2000 to 2011, *Computers & Education* **58**(2) (2012), 835-842.

[2] E. Alepis and M. Virvou, Automatic generation of emotions in tutoring agents for affective e-learning in medical education, *Expert Systems with Applications* **38**(8) (2011), 9840-9847.

[3] A. Anohina-Naumeca and J. Grundspeņķis, Evaluating students' concept maps in the concept map based intelligent knowledge assessment system. In: *Advances in Databases and Information Systems, Associated Workshops and Doctoral Consortium of the 13th East European Conference,* 7-10 September 2009, Riga, Latvia. Lecture Notes in Computer Science **5968**, Springer, Berlin, 2010, 8-15.

[4] C. Frasson and E. Aimeur, Designing a multi-strategic intelligent tutoring system for training in industry, *Computers in Industry* **37**(2) (1998), 153–167.

[5] N. L. Gage and D. C. Berliner (N. L. Geidžs and D. C. Berliners), *Pedagoģiskā psiholoģija/Educational Psychology*, Zvaigzne *ABC*, Riga, 1999.

[6] E. Gouli *et al.*, Exploiting COMPASS as a tool for teaching and learning, *Proceedings of the 3rd International Conference on Concept Mapping*, Tallinn, Estonia & Helsinki, Finland, **2** (2008), 383-390.

[7] J. Grundspeņķis, Usage experience and student feedback driven extension of functionality of concept map based intelligent knowledge assessment system, *Communication & Cognition* **43** (1-2) (2010), 13-32.

[8] K. Günel and R. Aşlıyan, Extracting learning concepts from educational texts in intelligent tutoring systems automatically, *Expert Systems with Applications* **37**(7) (2010), 5017-5022.

[9] F. Gutierrez and J. Atkinson, Adaptive feedback selection for intelligent tutoring systems, *Expert Systems with Applications* **38**(5) (2011), 6146-6152.

[10] Y. He, S. Ch. Hui, T. T. Quan, Automatic summary assessment for intelligent tutoring systems, *Computers & Education* **53**(3) (2009), 890-899.

[11] B. G. Johnson, F. Phillips, and L. G. Chase, An intelligent tutoring system for the accounting cycle: enhancing textbook homework with artificial intelligence, *Journal of Accounting Education* **27**(1) (2009), 30-39.

[12] B. Kort, R. Reilly, and R. W. Picard, An affective model of interplay between emotions and learning: reengineering educational pedagogy – building a learning companion. In: *Proceedings of IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges,* 6-8 August 2001, Massachusetts, USA. IEEE Computer Society, 2001, 43-48.

[13] A. Latham, K. Crockett, D. McLean, and B. Edmonds, A conversational intelligent tutoring system to automatically predict learning styles, *Computers & Education* **59**(1) (2012), 95-109.

[14] P. J. Muñoz-Merino, M. F. Molina, M. Muñoz-Organero, and C. D. Kloos, An adaptive and innovative question-driven competition-based intelligent tutoring system for learning, *Expert Systems with Applications* **39**(8) (2012), 6932-6948.

[15] V. L. Payne, O. Medvedeva, E. Legowski, M. Castine, E. Tseytlin, D. Jukic, and R. S. Crowley, Effect of a limited-enforcement intelligent tutoring system in dermatopathology on student errors, goals and solution paths, *Artificial Intelligence in Medicine* **47**(3) (2009), 175-197.

[16] P. Phobun and J. Vicheanpanya, Adaptive intelligent tutoring systems for e-learning systems, *Procedia - Social and Behavioral Sciences* **2**(2) (2010), 4064-4069.

[17] I. Roll, V. Aleven, B. M. McLaren, and K. R. Koedinger, Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system, *Learning and Instruction* **21**(2) (2011), 267-280.

[18] A. Sarrafzadeh, A. Samuel, D. Farhad, F. Chao, and B. Abbas, "How do you know that I don't understand?" A look at the future of intelligent tutoring systems, *Computers in Human Behavior* **24**(4) (2008), 1342-1363.

[19] S. Schiaffino, P. Garcia, and A. Amandi, eTeacher: Providing personalized assistance to e-learning students, *Computers & Education* **51**(4) (2008), 1744-1754.

[20] C. Woo Woo, *M. W. Evens, R. Freedman, M. Glass, L. S. Shim, Y. Zhang, Y. Zhou, and J. Michael,* An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning, *Artificial Intelligence in Medicine* **38**(1) (2006), 25-46.

[21] M. Yaghmaie and A. Bahreininejad, A context-aware adaptive learning system using agents, *Expert Systems with Application* **38**(4) (2011), 3280-3286.

Towards Automatic Structured Web Data Extraction System

Tomas GRIGALIS [a,1]

[a] *Vilnius Gediminas Technical University, Vilnius, Lithuania*

Abstract. Automatic extraction of structured data from web pages is one of the key challenges for the Web search engines to advance into the more expressive semantic level. Here we propose a novel data extraction method, called ClustVX. It exploits visual as well as structural features of web page elements to group them into semantically similar clusters. Resulting clusters reflect the page structure and are used to derive data extraction rules. The preliminary evaluation results of ClustVX system on three public benchmark datasets demonstrate a high efficiency and indicate a need for a much bigger up-to-date benchmark data set that reflects contemporary WEB 2.0 web pages.

Keywords. Information extraction, structured web data, deep web

Motivation and Research Questions

For the Web search engines to advance into a more expressive semantic level, we need tools that could extract the information from the Web and represent it in a machine readable format such as RDF [1]. Information extraction at Web scale is the first step in persuing this goal. However, current algorithmic approaches often fail to achieve satisfactory performance in real-world application scenarios due to abundant structurally complicated WEB 2.0 pages.

In this work we address the problem of automatic information extraction from structured Web data, such as lists of products in online stores. We propose a novel approach, called ClustVX, which is fully automatic, scalable, and domain independent.

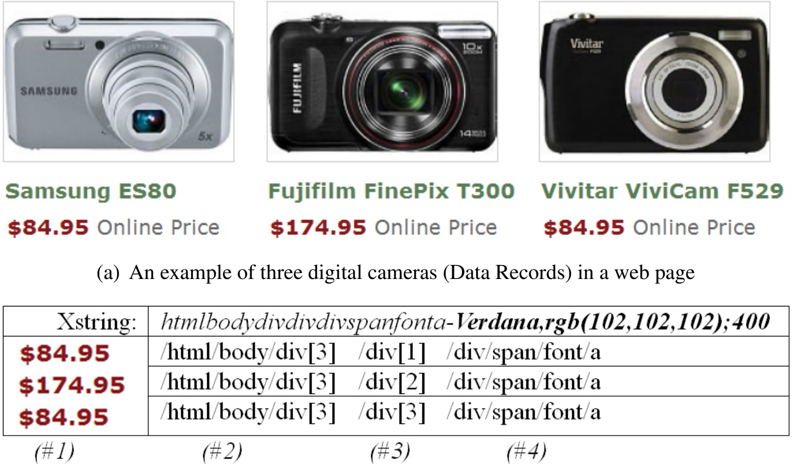
ClustVX is based on two fundamental observations. First, vast amount of information on the Web is presented using fixed templates and filled with data from underlying databases. For example, Fig. 1(a) shows three Data Records (DRs) representing information about three digital cameras in an online store. The three DRs are listed according to some unknown to us style template and the information comes from a database. This also means, that each DR has almost the same Xpath (tag path from root node in HTML tree to particular web page element), where only a few node numbers differs.

Second, although the templates and underlying data differ from site to site, humans understand it easily by analyzing repeating visual patterns on a given Web page. We hypothesize, that the data which has the same semantic meaning is visualized using the same style. Therefore humans, viewing such a web page, are able to comprehend its

[1] PhD student. Email: tomas.grigalis@vgtu.lt; Supervisor: Antanas Cenys, Email: antanas.cenys@vgtu.lt

unique structure quickly and effortlessly and distinguish items photos, titles, prices and etc. For example in Fig. 1(a) prices are brown red and bold, title is green and bold, text ”Online Price” is grey.

ClustVX exploits both of these two observations by representing each web page element with a combination of its Xpath and visual features such as font, color and etc. For each visible web page element we encode this combination into the string called Xstring. Clustering Xstrings allows us to identify visually similar elements, which are located in the same region of a web page and in turn have same semantic meaning. See Fig. 1(b) where price elements are clustered together according to their Xstring. Subsequent data extraction leads to a machine readable structured data. The result of this extraction is shown in Fig. 1(c). Our preliminary evaluation on three public datasets demonstrate that the new method is able to consistently achieve high recall and precision in extracting structured data from given web pages.



(a) An example of three digital cameras (Data Records) in a web page

Xstring:	*htmlbodydivdivdivspanfonta-**Verdana,rgb(102,102,102);400***
$84.95	/html/body/div[3] /div[1] /div/span/font/a
$174.95	/html/body/div[3] /div[2] /div/span/font/a
$84.95	/html/body/div[3] /div[3] /div/span/font/a

(#1) (#2) (#3) (#4)

(b) A cluster with visually similar price elements

Image 1	Samsung ES80	$84.95	Online Price
Image 2	Fujifilm FinePix T300	$174.95	Online Price
Image 3	Vivitar ViviCam F529	$84.95	Online Price

(c) Desired extraction result

Figure 1. An example of information extraction using ClustVX

In the following we present a brief review of the current related research work and then, in Sec. 2 we outline the ClustVX system. We present experimental results in Sec. 3 and, finally, outline the necessary future research directions and further aspects of experimental evaluation in Sec. 4.

1. Related Work

Data extraction systems can be broadly divided into supervised and unsupervised categories. Supervised learning approaches require some manual human effort to derive the extraction rules, while automated data extraction systems work automatically and need no manual intervention to extract data.

In this work we focus on the latter as we believe that only fully automatic systems can be applied for web-scale data extraction. Our proposed ClustVX system belongs to this category.

One widely adopted technique to automatically detect and extract DRs is to search for repetitive patterns in HTML source code by calculating the similarity of HTML tree nodes. Variations of simple tree matching algorithm [10] are employed for this task [3,8]. However, this technique finds it difficult to deal with structural irregularities amongst DRs , such as lists inside DRs [11].

Contrary to the above recent system VIDE [4] tries to not tie itself to HTML tree at all and instead depends purely on visual features of a web page. It builds a visual containment tree of a web page using patented VIPS [7] algorithm and then uses it instead of HTML tree. However if there are some unloaded images or missing style information in a web page VIPS may fail to build correct visual containment tree which leads to data extraction problems [4].

Combining previous two approaches ViNTs [5] and DEPTA [3] systems try to exploit visual features of web pages to aid structural based data extraction process. However, ViNTs system do not extract data items, it just segment DRs, and evaluation of DEPTA demonstrated, that it cannot handle contemporary pages efficiently [11].

Systems like TextRunner [6] try to extract entities and their relationships from web pages using natural language processing and machine learning approaches, but those techniques usually work on regular text and are not suitable for detecting repetitive patterns in web pages. By contrast, WebTables [2] extracts entities from structured web tables enclosed in $<$ `table` $>$ HTML tags. However, it would miss structured data presented in other form of html tags.

In summary, none of the systems can properly handle visual and structural features of web page to effectively extract structured web data. The ClustVX system proposed in this work fully exploits visual and structural information and achieves promising results.

2. The ClustVX Approach

The ClustVX processes a given Web page in the following steps:

1. Preprocessing the page. In this stage the web page is cleaned from all HTML text formatting tags, such as $<$ `b` $>$, $<$ `em` $>$, which appear in the middle of a text and may hinder the clustering process. Visual features of web page elements acquired from browser's API are embedded into HTML tags for processing in the next step.
2. Generating Xstring representation of each HTML element. Each visible web page elements is represented by a Xstring, by which the elements are later clustered. As we see in Fig. 1(b) Xstring consists of a) tag names from Xpath b) visual features of that element (font style, color, weight, etc.). Structural features (string of tag names) identifies position in HTML document. Visual features enhance understanding of semantic similarity between web page elements.
3. Clustering of web page elements. All visible web page elements are clustered according to their Xstring. Resulting clusters contain only semantically similar web page elements. In Fig. 1(b) at (#1) we see a cluster of price elements.

Table 1. Publicly available benchmark data sets for structured web data extraction

Data Set	TBDW [9]	ViNTs-2 [5]	Alvarez [8]
Sites	51	102	200
Pages per site	5	11	1
AVG records per page	21	24	18
Total records (1st page per site)	1052	2489	3557

4. Extraction of structured data. By analyzing Xpath of clustered visually similar web page elements, extraction rules are induced and data is extracted. In Fig. 1(b) at (#2) is a Xpath of the region in a page where Data Records are located. Each Data Record is enclosed in a DIV tag (#3). The final path of price elements inside a Data Record is (#4).

3. Research Methodology

To evaluate ClustVX approach we use the following three publicly available benchmark datasets containing in total of 7098 data records: 1) TBDW Ver. 1.02 [9], 2) ViNTs dataset 2 [5], 3) M. Alvarez et al. [8]. See Tab. 1 for details. These data sets contain search result pages generated from databases. Following the works of other authors [8, 11,4,3,5] in structured data extraction we use three evaluation metrics which come from information retrieval field: precision, recall and F-score.

The positive preliminary results showing that ClustVX can achieve higher than 0.98 F-Score encourage further development and evaluation of ClustVX on real-world data. We see a must to have a data set containing thousands of pages from different web sites. To create such a huge data set we are planning to exploit the power of crowdsourcing by the help of Amazon Mechanical Turk service [12]. The sercvice lets to present simple human intelligence requiring tasks, such as labeling data or telling if extraction was successful, to thousands of voluntary workers, which are paid on per hour or per task basis.

4. Conclusions and Research Directions

In this paper we presented ClustVX system, which, by exploiting visual and structural features of web page elements, extracts structured data. The preliminary evaluation of ClustVX on three publicly available benchmark data sets demonstrated, that our method can achieve very high quality in terms of precision and recall. Our future work will be concentrated on creating a new huge benchmark data set, dealing with extremely malformed HTML source code and comparing ClustVX system to competing approaches.

References

[1] Weikum, G., Theobald, M.: From information to knowledge: harvesting entities and relationships from web sources. In: *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM (2010), 65–76.

[2] Cafarella, M., Halevy, A., Wang, D., Wu, E., Zhang, Y.: Webtables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment* **1**(1) (2008), 538–549.

[3] Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: *Proceedings of the 14th international conference onWorld WideWeb*, ACM (2005) 76–85.

[4] Liu, W., Meng, X., Meng, W.: Vide: A vision-based approach for deep web data extraction. *IEEE Transactions on Knowledge and Data Engineering* **22**(3) (2010), 447–460.

[5] Zhao, H., Meng, W., Wu, Z., Raghavan, V., Yu, C.: Fully automatic wrapper generation for search engines. In: *Proceedings of the 14th international conference on World Wide Web*, ACM (2005), 66–75.

[6] Banko, M., Cafarella, M., Soderland, S., Broadhead, M., Etzioni, O.: *Open information extraction for the web*, University of Washington (2009).

[7] Cai, D., Yu, S., Wen, J., Ma, W.: Vips: a visionbased page segmentation algorithm. Tech. rep., Microsoft Technical Report, MSR-TR-2003-79 (2003).

[8] Alvarez, M., Pan, A., Raposo, J., Bellas, F., Cacheda, F.: Extracting lists of data records from semi-structured web pages. *Data and Knowledge Engineering* **64**(2) (2008), 491–509.

[9] Yamada, Y., Craswell, N., Nakatoh, T., Hirokawa, S.: Testbed for information extraction from deep web. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers and posters*, ACM (2004), 346–347.

[10] Yang, W.: Identifying syntactic differences between two programs. *Software: Practice and Experience* **21**(7) (1991), 739–755.

[11] Jindal, N., Liu, B.: A generalized tree matching algorithm considering nested lists for web data extraction. In: *The SIAM International Conference on Data Mining* (2010), 930–941.

[12] Alonso, O., Rose, D., Stewart, B.: Crowdsourcing for relevance evaluation. In: *ACM SIGIR Forum*, ACM, Vol. 42 (2008), 9–15.

A Comparison of SOA Methodologies Analysis & Design Phases

Sandra SVANIDZAITĖ
Institute of Mathematics and Informatics, Vilnius University

Abstract. Service oriented computing is a new software engineering paradigm that represents a shift in software engineering and raises the abstraction level by grouping common business process functionality and exposing it as a service. SOA allows a rapid and low-cost application development through service composition. Existing widely used methodologies designed to support object-oriented development such as RUP or agile cannot be reused for SOA without any adaptation. As a consequence, new methodologies that address all the principles and patterns of SOA are required to ensure effective SOA application development. This paper aims to present a state-of-the-art of the most widely known SOA methodologies describing their solutions proposed for SOA analysis & design phases. The characteristics according to which these methodologies are compared are discussed. The results of comparison are provided.

Keywords. SOA, SOA analysis, SOA design, SOMA, SOAF

Introduction

Most SOA methodologies propose to divide SOA development lifecycle into six phases: Service-oriented analysis, Service-oriented design, Service development/construction, Service testing, Service deployment/transition, Service administration/management. The first two phases are the most important ones because the success of SOA development mainly depends on them. Technology and standards, such as BPM, BPEL, WSDL, EA, OOAD are important to develop SOAs, but it has been widely recognized that they are not sufficient on their own. Just by applying a Web service layer on top of legacy applications or components does not guarantee true SOA properties, such as business alignment, flexibility, loose coupling, and reusability. Instead, a systematic and comprehensive SOA analysis & design methodology is required [1]. A number of SOA methodologies such as IBM RUP/SOMA, SOAF, SOUP, methodology by Thomas Erl and methodology by Michael Papazoglou has been proposed to ensure successful SOA development. A number of SOA methodology surveys have already been performed but they treat them from a general point of view without providing any in-depth analysis of properties of these methodologies aiming at SOA analysis & design phases. This paper contributes to outlining the drawbacks and benefits of proposed SOA methodologies and focuses on SOA analysis & design phases by providing in-depth analysis and a comparison according to characteristics specified.

1. Characteristics of SOA Methodologies Analysis & Design Phases

In order to evaluate analysis & design phases in SOA methodologies we have defined characteristics that will be used to perform a comparison. The characteristics proposed for evaluation are as follows [1, 2, 3, 4, 5]:

SOA analysis & design strategy: Three strategies (top-down, bottom-up and meet-in-the-middle) exist in SOA development, each varying in the amount of up-front analysis of the business domain and the dependencies on legacy systems.

SOA analysis & design coverage: Service-oriented analysis and design phases of SOA methodologies that will be analyzed and compared can be divided into five main activities that are further refined into steps. These steps are used for evaluation of SOA analysis & design coverage.

Main activities of SOA analysis & design phases:

- **Target Organization‘s Business analysis**. The aim of this step is to identify: organization‘s objectives, business goals and KPIs for their accomplishment. Also used technology, applications and people skills, common business terms vocabulary, business rules, business actors and main business use cases are defined. The step results in the creation of “as-is“ and “to-be“ business models.
- **SOA project planning.** The aim of this step is to formulate the vision and the scope of SOA project, select SOA delivery strategy (create services from scratch, create services from existing software components, buy services from third party providers), create project plan and accomplish financial analysis.
- **Service Identification.** The aim of this step is to identificate candidate services. All functional and non-functional requirements for SOA development are gathered. Created “to-be” business model is decomposed into business domains. After that, service candidates, their initial specifications, communication and initial dependencies are defined. Existing applications are analyzed in order to find which software components can be reused in SOA development.
- **Service Analysis and Specification.** The aim of this step is to select which candidate services will be developed and to create detailed service specifications for development. Services are grouped by their functionality into: business entity, application and business process services. Business process specifications that will group the services are created.
- **Service Realization Decisions.** The aim of this step is to document service realization decisions, to allocate service components to layers and to accomplish technical feasibility exploration.

Degree of prescription: SOA methodologies vary from the most prescriptive ones to the less descriptive ones. The degree of prescription is evaluated depending on the number of parameters provided in process description. Available parameters are: phases, activities, steps and inputs, outputs for each step.

Adoption of existing techniques and notation: Most of SOA methodologies are based on techniques such as OOAD, CBM, BPM, EA and notations such as UML and BPMN, while the others do not address specific techniques and notations and let the user to decide what techniques and notations are appropriate in a concrete situation, making the methodology harder to understand and to use for inexperienced users.

2. Analysis of Existing SOA Methodologies

IBM RUP/SOMA [6] is an integrated methodology developed by IBM in a will to bring unique aspects of SOMA to RUP. However, because SOMA is a proprietary methodology of IBM, its full specification is not available.

Methodology consists of four phases: *business transformation analysis, identification, specification, and realization of services*. Talking about SOA analysis & design all these phases are of great importance. However IBM RUP/SOMA does not cover the deployment and administration of services. The first phase *Business Transformation Analysis* can be mapped to *Inception* phase from classical RUP methodology. This phase is an optional one and can be omitted if organization‘s full business analysis and transformation is not performed. It aims to describe current "as-is" organization business process, to understand problem areas and improvement potentials as well as any information on external issues such as competitors or trends in the market. *Business Transformation Analysis* comprises such activities as: assessment of target organization and its objectives, identification of business goals and KPIs, definition of common business vocabulary and business rules, definition of business actors and main use cases, analysis of business architecture. The second phase *Service Identification* can be mapped to *Elaboration* phase from classical RUP and aims to identificate candidate services. Service Identification comprises such activities as: *Domain Decomposition, Goal-Service Modeling and Existing Asset Analysis*. The third phase *Service Specification* can be mapped to *Elaboration* phase from classical RUP and focuses on the selection of candidate services that will be developed. Service Specification phase comprises such activities as: *Service Specification, Subsystem Analysis and Component Specification*. The fourth phase *Service Realization* can be mapped to *Construction* phase from classical RUP and is focused on completion of component design for component implementation. Service Realization comprises such activities as: *Documentation of Service Realization Decisions, Allocation of Service Components to Layers*.

Service Oriented Architecture Framework (SOAF) [7] methodology consists of five main phases: *information elicitation, service identification, service definition, service realization, roadmap and planning*. The aim of SOAF is to ease the service identification, definition and realization activities by combining a top-down modeling of an existing business process with a bottom-up analysis of existing applications. The first phase *Information Elicitation* aims to define the scope and constraints of existing business process and used technology. Current business "as-is" model is created and "to-be" business model is defined. Candidate services are identified that will automate "to-be" business model. Non-functional requirements (NFRs) and Business Level Agreements (BLAs) should be also defined, categorized and prioritized. Process-to-Application Mapping (PAM) is performed that examines existing software assets in order to discover SOA candidate application functionality. *Service Identification* phase aims to define an optimal set of services. *Service realization* phase aims to define transformation strategies that will be used for transition from the legacy application architecture to the future application architecture by reusing, developing and buying third party services. The *roadmap and planning* phase purposes a detailed planning of transformation and identifies business and technical risks.

Methodology by Papazoglou [1]. In the paper, Papazoglou et al provide a SOA development methodology that covers a full SOA lifecycle. It is partly based on well-established development methodologies as RUP, Component-based Development and BPM [5]. The methodology is based on iterative and incremental process and comprises one preparatory - *Planning* and eight main phases: *Service Analysis, Service Design, Service Construction, Service Test, Service Provisioning, Service Deployment, Service Execution and Service Monitoring.* Talking about SOA analysis & design only the *Planning, Service Analysis* and *Service Design* phases are important. The *Planning* phase is a preparatory one. Activities in this phase include analysis of business needs and review of current technology landscape, financial analysis of the project and a creation of SOA development plan. The aim of *Service-oriented analysis* phase is to elicit the requirements for SOA application. Business analysts create an "as-is" business process model that allows stakeholders to understand a portfolio of available services and business processes. The phase results in creation of the "to-be" business process model that will be implemented in SOA solution. Analysis phase consists of four main activities: *process identification, process scoping, business gap analysis and process realization*. *Service Design phase* aims to transform business processes and services descriptions to well-documented service interfaces and service compositions. Design phase consists of two activities: *Specification of Services and Specification of Business Processes*.

Methodology by Thomas Erl [2], [3]. This methodology is a step by step guide through the two main phases: service-oriented analysis and design. *Service-oriented analysis* comprises three main steps: *define business requirements, identify existing automation systems and model candidate services* and can be divided in two main parts: the first part in which business requirements are defined and the second part in which service candidates are modeled. The first part of the phase includes reviewing business goals and objectives, analyzing potential changes to existing applications in a will to find which processes and application components can be used in a future SOA application development. Business analysts prepare "as-is" process model which states the current situation and allows stakeholders to understand which business processes are already in place and which has to be introduced and automated, which application components can be reused. Service-oriented analysis results in the preparation of "to-be" process model that an SOA application will implement. The second part of service-oriented analysis is a *service modeling* sub-process by which service candidates are identified and modeled. Service modeling sub-process results in the creation of such artifacts as: *conceptual service candidates, service capability candidates and service composition candidates*.

Service-oriented Unified Process [8] or SOUP is a hybrid software engineering methodology that is targeted at SOA projects. As the name suggests this methodology is primarily based on the Rational Unified Process. Its lifecycle consists of six phases: *Incept, Define, Design, Construct, Deploy* and *Support*. SOUP methodology can be used in two slightly different variations: one adopting RUP for initial SOA projects and other adopting a mix of RUP and XP for the maintenance of existing SOA applications. When talking about SOA analysis only the first three phases *Incept, Define* and *Design* of this methodology are important. The first *Incept* phase aims to understand the business needs for SOA development and how SOA fits within the organization. The objective of this phase is to decide whether SOA project is profitable by evaluating project scope and risks or not. Incept phase comprises such activities as*: Formulation of the vision and of the scope of the system, Definition of SOA strategy, Return-on-Investment (ROI) analysis accomplishment* and *Creation of Communication Plan*. The second *Define* phase is the most critical phase in SOA project. It aims to define the requirements and develop use cases. The objectives of this phase are: 1) to fully understand business processes affected, 2) to collect, define and analyze functional and non-functional requirements by using a formal requirements-gathering and management process like RUP, 3) to design support and governance model which explains how organization will support SOA, 4) to prepare a realistic project plan, 5) to define a technical infrastructure that is required to support entire SOA. The third *Design* phase aims to translate use case realizations and SOA architecture into detailed design documents. The objectives of this phase are: 1) to create detailed design document and data base model that explain the structure of the services, 2) to structure development process by defining the technology, coding standards and etc.

3. Comparison of SOA Methodologies

Analyzed and described in 2 section SOA methodologies were compared using characteristics described in 1 section by outlining main differences, benefits and drawbacks. Detailed comparisons are not included in the paper due to the space limits. The comparison resulted in a number of insights:

- The most prescriptive SOA methodology is IBM RUP/SOMA which is a proprietary one and widely used in industrial projects. It supports *meet-in-the-middle* SOA analysis & design strategy, covers all SOA analysis & design activities. It also has the best degree of prescription, because it provides activities, steps, inputs and outputs description for each phase. It adopts such existing techniques and notation as: BPM, UML, BPEL, WSDL, WS-BPEL.
- A methodology by Thomas Erl does not provide detailed descriptions how to start the SOA project, how to perform organization‘s business analysis, how to formulate the vision and the scope of the project, but, it provides detailed service-oriented analysis & design phases descriptions meaning that it cannot be used from the start of the project but it can be used in conjunction with other methodology that provides detailed recommendations how to initiate SOA project. It supports *top-down* SOA analysis & design strategy, has a good degree of prescription and also adopts such existing techniques as: BPM, WSDL, WS-BPEL, WS-* specifications.
- SOUP methodology is still only in its first steps and is not mature enough to assure successful SOA development because it lacks prescription: phases, activities, artifacts, process workers and their roles are not defined clearly. It supports *meet-in-the-middle* SOA analysis & design strategy, but it does not cover some of the SOA analysis & design activities. SOUP methodology lacks adoption of existing notations such as UML and BPMN that are used in service-oriented analysis and design.
- SOAF methodology supports *meet-in-the-middle* SOA analysis & design strategy, but it does not cover some of the SOA analysis & design activities, lacks prescription and adoption of existing techniques and notations to assure successful SOA development.
- Methodology by Papazoglou supports *meet-in-the-middle* SOA analysis & design strategy, adopts such techniques and notations as: CBD, BPM, BPMN, WSDL, BPEL, UML. It provides detailed recommendations for Service

Design and Specification, but as a methodology for SOA analysis & design it lacks prescription. It does not refine activities in concrete steps and does not provide inputs and outputs for them.

4. Conclusions

The aim of this paper was to compare the most widely known and popular SOA development methodologies by providing an in-depth analysis of Service-oriented analysis and design phases. In this paper we analyzed and compared the following SOA methodologies: IBM RUP/SOMA, SOAF, methodology by Thomas Erl, methodology by Papazoglou and SOUP. The research showed that: analyzed SOA methodologies vary in a degree of prescription from the most prescriptive ones, to the less prescriptive ones letting the user to tailor and to adapt the methodology to concrete project‘s scope. In addition to this, most of analyzed SOA methodologies are built upon and incorporate existing and proven techniques, notations such as OOAD, CBD, BPM, WSDL, BPEL, UML, meaning that earlier used approaches are still applicable and new ones for SOA development are offered, but new method for organizing the process of SOA development is lacking. Most of analyzed SOA methodologies propose meet-in-the-middle strategy for Service-oriented analysis, meaning that most of SOA projects do not start in an empty place and most of them are targeted to change legacy systems. Service-oriented analysis and design phases in each methodology result in similar list of key deliverables, although each methodology offers a slight different but at some activities overlapping approach to achieve them.

In the conclusion, we can say that much is already done in this area, but there is still a lack of mature, descriptive, validated in proof-of-concept case studies, non-proprietary SOA methodology.

References

[1] M. P. Papazoglou and W.-J. van den Heuvel, Service-oriented design and development methodology, *International Journal of Web Engineering and Technology* **2**(4) (2006), 412-442.

[2] T. Erl, *Service-Oriented Architecture: Concepts, Technology and Design*. Prentice Hall PTR, 2005.

[3] T. Erl, *SOA Principles of Service Design*. Prentice Hall PTR, 2008.

[4] H. M Shirazi, N. Fareghzadeh, and A. Seyyedi, A combinational approach to service identification in SOA, *Journal of Applied Sciences Research* **5**(10) (2009), 1390-1397.

[5] E. Ramollari, D. Dranidis, and A. J. H. Simons, A survey of service oriented development methodologies. In: *Proceedings of the 2nd European Young Researchers Workshop on Service Oriented Computing*, Leicester, UK, June 2007.

[6] *Introduction to RUP*, IBM Corp. [cited April 2012]. Available from: http://www.michael-richardson.com/rup_classic/#core.base_rup/guidances/supportingmaterials/introduction_to_rup_36B63436.html.

[7] A. Erradi, S. Anand, and N. Kulkarni, SOAF: an architectural framework for service definition and realization. In: *IEEE International Conference on Services Computing (SCC 2006),* 18-22 September 2006, Chicago, Illinois, USA. IEEE Computer Society, 2006, 151-158.

[8] K. Mittal, *Service Oriented Unified Process*. [cited April 2012]. Available from: http://www.kunalmittal.com/html/soup.html.

Operationalization of Norms in Aircraft Approach/Departure Decision Support

Laura SAVIČIENĖ
Faculty of Mathematics and Informatics, Vilnius University, Lithuania

Abstract. This work is focused on norm operationalization in aviation domain. The investigated paradigm can be described as: from legal norms to technical rules in the artifact. Normative requirements (norms) for the aircraft trajectories are extracted from the flight rules and airport procedures, and operationalized in a decision support system (DSS). The decision support is based on evaluation of risk to violate the normative requirement. The following risks are modeled: trajectories' conformance with the flight rules, safe distance between aircraft, wake turbulence separation and avoidance of volcanic ash. The DSS is for the air traffic controller (not pilot) and must respond in real time. It provides surveillance, evaluates and recommends, whereas the human controller takes a decision.

Keywords. Air traffic control, instrument approach procedure, SKY-Scanner, real-time decision support, norm operationalization, risk model

Introduction

This research is focused on the operationalization of normative rules in aviation domain (air traffic control, ATC). A proposed paradigm can be called “from legal norms to technical rules in the artifact”. Normative requirements are extracted from the flight rules, maps and approach/departure procedure charts. An example of a normative rule is “Keep 3 degrees descent angle while landing and hold restrictions of the altitude and geography depicted in the aerodrome chart”.

Normative rules are modeled in order to provide decision support in terms of norm violation risk. A decision support system (DSS) provides surveillance, evaluates and recommends, whereas the human controller takes a decision. The final decision is done by human controller. This approach accords with SESAR (Single European Sky Air Traffic Management Research) target concept, which states that humans should constitute the core of the future air traffic management (ATM) operations [1].

The decision support is based on lidar (laser radar – LIght Detection And Ranging) and radar data fusion. It relies on the assumption that the precise aircraft position data (with error margin of meters, not hundreds of meters) from the lidar will facilitate detection of risks that are not possible to detect using only radar data.

The research goal is to develop a conception for operationalization of the aircraft approach/departure norms in a decision support system, taking into consideration the use of lidar for aircraft tracking. The goal is broken down into these tasks: (1) modeling norm violation risk in the airport traffic zone (ATZ), (2) modeling radar and lidar data fusion, and (3) development of a prototype decision support system.

1. Domain Analysis and Related Works

The norm operationalization is investigated in the context of the ATM paradigm developed in the EU FP6 SKY-Scanner project[1]: expanding surveillance and the ATC control to the approach/departure phases by using radar and lidar data fusion and decision support in terms of norm violation risk. Only norms that can be checked using the lidar-radar fused data (position and speed) are examined and included in the operationalization conception.

The project was aimed at developing a laser system to detect and track aircraft up to at least 6 nautical miles (NM) from the aerodrome traffic zone (ATZ) barycenter [2]. The project objectives include aircraft collision probability model (ACPM) based on radar data and laser tracking data fusion and a prototype decision support system (Figure 1) for aircraft approach and departure [3]. The current research builds on the constraint models developed by the SKY-Scanner project, and aims to abstract them into a unified norm operationalization conception, also further refining the DSS prototype and visualization models.

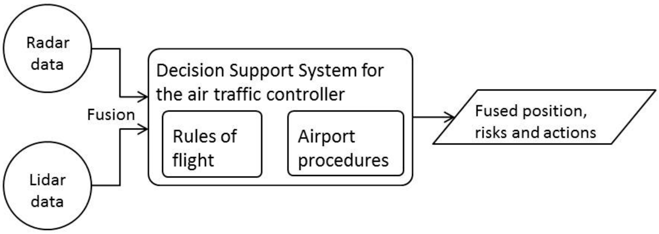

Radar data
Fusion
Lidar data
Decision Support System for the air traffic controller
Rules of flight
Airport procedures
Fused position, risks and actions

Figure 1. DSS for aircraft approach and departure

Several important assumptions stem from the SKY-Scanner project and thus form the boundaries of the current research:

1) ATC activities require a real-time response from the DSS. A study of time-critical decision support models provided in [4] concludes that the naturalistic decision support approach should be used and highlights the need to filter out the most important information for the user.

2) The emphasis is on informing the controller, who then makes a final decision on the actions. This accords with the results of studies of human-automation interaction: high levels of automation are not advisable in systems dealing with dynamic environments with many external and changing constraints [5].

Normative rules for aircraft approach/departure from International Civil Aviation Organization (ICAO) flight rule documents [6, 7, 8, 9] are grouped into four categories: ATC separation rules, airport procedures, wake turbulence separation rules, and volcanic ash rules. Each airport has a different set of approach/departure procedures. Approach/departure procedure constitutes a complex object and contains a number of interrelated norms that define the ought-to-be trajectory with additional constraints.

Current aviation-related decision support systems do not model norms comprehensively, but there is some research in that direction. One type of decision support – Conflict Detection and Resolution (CD&R) systems. The structure of the

[1] "Development of an Innovative LIDAR Technology for New Generation ATM Paradigms" (SKY-Scanner), 2007-2010, http://www.sky-scanner.it/

CD&R process [10] is designed for the aircraft separation conflicts, but can be expanded to cover other normative rules. Conformance alerting philosophy is suitable for the approach/departure norm supervising scenario: alert is issued when the aircraft is close to violating the norm.

2D visualizations in the ATM domain are no longer sufficient, and the modern 3D visualizations have drawbacks [3, 11]. By augmenting the 3D screens with auxiliary 2D elements it is possible to visualize the ought-to-be trajectory requirements: a relationship between horizontal position, distance and altitude.

2. Norm Conceptualization and Risk Modeling

The approach/departure decision support focuses on detecting violations of the flight rules for the aircraft. We conceptualize each norm as a triplet of a norm factor, norm pattern, and the expected value. Norm factor represents a quantitative trajectory attribute of one or several aircraft. Only factors that can be computed from the DSS input data are considered in this conception. Expected value, v_N, is the value defined in the text of the normative requirement. Norm pattern (‘$\leq v_N$’, ‘$\geq v_N$’, or ‘$=v_N$’) explicates how to interpret the expected value. For example, norm pattern ‘$\geq v_N$’ means the actual value of the factor should be greater than the expected value.

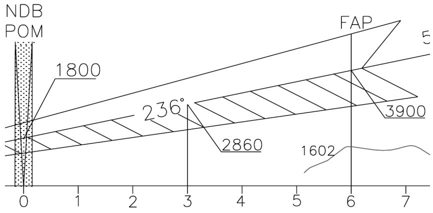


Figure 2. Examples of norms in the approach chart [12]

Patterns ‘$\leq v_N$’ and ‘$\geq v_N$’ constitute limit-based norms, and pattern ‘$=v_N$’ – deviation-based norms. Example of the limit-based norm (Figure 2): “height minimum is 3900 ft. at 6 nautical miles from distance measurement equipment (DME)”. An example of the deviation-based norm could be the track (the direction that the aircraft should follow), which is expressed in degrees from North, e.g. 236° [12].

The defined operationalization structure translates each norm into a risk definition in the DSS. The use of discrete risk levels abstracts from unnecessary details. In the DSS risk levels are defined based on the likelihood of violating the norm. A separate risk definition is formulated for each normative requirement. An individual risk evaluation maps the observed factor value to a discrete scale of risk levels.

The L-level risk concept is characterized by five elements (Figure 3): (1) risk factor (e.g. ‘altitude’ or ‘indicated airspeed’); (2) risk type (‘limit’ or ‘deviation’); (3) the norm pattern (‘$\geq v_N$’, ‘$\leq v_N$’, ‘$=v_N$’); (4) expected value of the factor; (5) a set of thresholds for risk levels. If the risk type is ‘limit’, a set of thresholds consists of L-1 constants, defined in the terms of factor measurement units. If risk type is ‘deviation’, a set of thresholds consists of L-1 pairs of constants, defining allowable deviation levels.

For convenience of visual representation of the risk definition, a piece-wise linear risk-magnitude function is used, which maps the observed factor value to a number from the interval [0, 1]. Zero means the lowest risk, 1 means the highest risk level, and values in the interval (0, 1) mean intermediate risk levels.

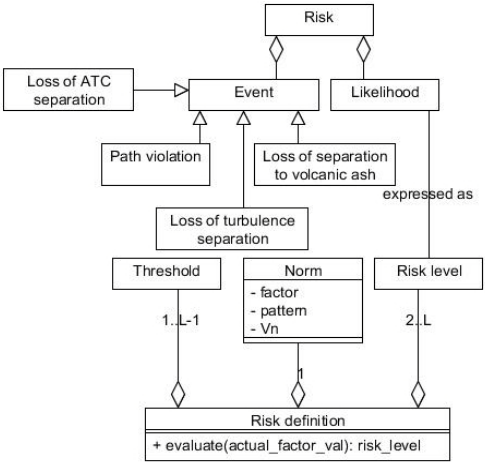


Figure 3. Risk definition

As an example we consider the norm "altitude 3900 ft at 6 DME" (Figure 2). In deviation-based risk evaluation we consider not the expected value itself, but the expected deviation (d_N = 0). There are 4 risk levels and 3 pairs of thresholds. The corresponding risk definition is: (1) factor: 'altitude'; (2) type: 'deviation'; (3) pattern: '=v_N'; (4) expected value: 3900 ft at 6 DME (deviation 0); (5) thresholds: d_{n0} = -0.5, d_{p0} = 2, d_{n1} = -1, d_{p1} = 3.5, d_{n2} = -1.5, d_{p2} = 5; (see Figure 4) The threshold values in this example are chosen only for demonstration purposes.

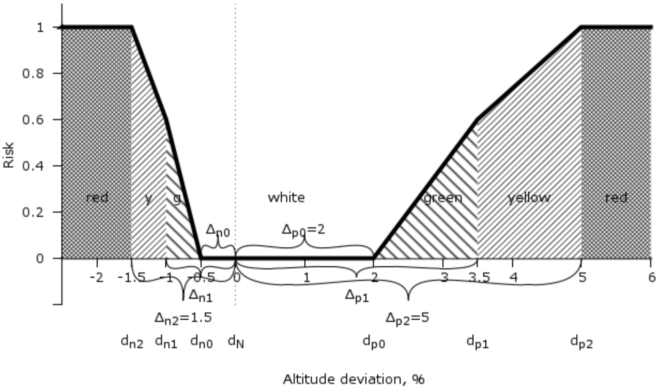


Figure 4. Altitude violation risk-magnitude function

Each risk is represented in a separate indicator on the DSS control panel (Figure 5). The risk level is shown with color and the number of colored slots on the indicator.

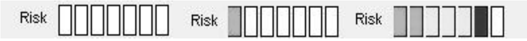


Figure 5. Risk indicators

3. DSS Prototype

The DSS prototype embodies the norm operationalization conception proposed in the previous chapter. It illustrates the modeling of several norms for the approaching aircraft, and provides a real-time simulation of the suggested decision support scenario.

Advanced ATC visualization ideas are adapted to context of approach/departure decision support. The DSS prototype provides a laboratory implementation, which advances technology readiness level (TRL [13]) 1-2 ideas to level 3. The aim was to visualize airport procedure requirements (the ought-to-be trajectory), so that the controller could visually estimate possible violation without looking at the control panel. Two visualizations are developed: 2D-in-3D prototype and pure-3D prototype. Both visualizations embed auxiliary 2D elements into the main 3D view of the observed airspace.

2D-in-3D prototype uses generalized terrain model and embeds 2D semi-transparent projection walls (Figure 6). Aircraft are represented with spheres. The ought-to-be trajectory is projected on the wall (white line), as well as the aircraft position (black dot). If the dot is not on the line, there is a path violation.

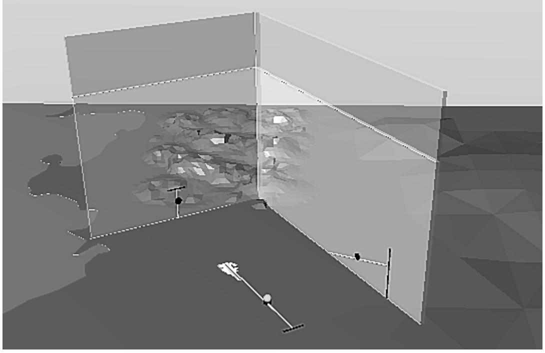


Figure 6. 2D-in-3D visualization

The pure-3D prototype uses photographic terrain (high resolution satellite images) and represent the aircraft as full 3D models (Figure 7). 2D rings enclose the ought-to-be trajectory (plus some allowable deviation). The violation is detected when the aircraft indicator is outside the rings. This approach is less strict that the 2D-in-3D.

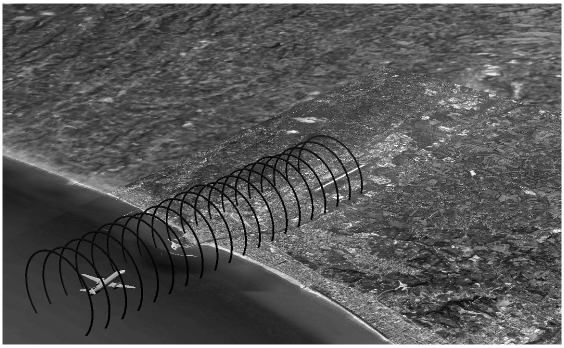


Figure 7. Pure-3D visualization

Human operator needs are satisfied in the following way: 3D display improves situation awareness as the airport environment is depicted with essential terrain obstacles; 2D elements (walls, ring) that relate the aircraft position to the airport procedures reduce cognitive workload.

4. Conclusions

The following conclusions are drawn:

1. The proposed norm operationalization conception enables to represent a subset of aircraft approach/departure normative rules in a decision support system for the air traffic controller. The referred subset is defined as the norms concerning aircraft trajectories, or simply, geometrical norms.
2. The prototype decision support system provides an integrated solution to facilitating the controller: risk indicators automate detection of the possible norm violations, 2D-in-3D visualizations help comprehend conformance to the approach/departure procedure. The controller's work improvement is not quantifiable because the research addresses the needs that will only become relevant in the future.
3. Analysis of the prototype development process demonstrates that the following steps are needed to operationalize a norm: (a) setting up risk levels and colors, (b) creating risk definitions (consisting of norm factor, expected value, type, pattern and a set of thresholds), and (c) setting up risk indicators. The process cannot be fully automated, as each norm factor is unique, and analysis has to be performed for each new kind of norm to be operationalized.

References

[1] *The ATM Target Concept*, SESAR Consortium, Toulouse, France, Sept. 2007.

[2] M. Salerno, D. Rondinella, M. V. Crispino, G. Costantini, M. Carota, and D. Casali, SKY-Scanner: a new paradigm for air traffic management, *International Journal Of Circuits, Systems And Signal Processing* **2**(2) (2008), 131-139.

[3] S. Rozzi, A. Boccalatte, P. Amaldi, B. Fields, M. Loomes, and W. Wong, *Innovation and Consolidation Report*, Middlesex University, London, UK, 2007.

[4] K. Lapin, SKY-Scanner: time-critical decision support system surveilling aircraft landing and take-off. In: *Proceedings of 9th Innovative Research Workshop & Exhibition*, Brétigny-sur-Orge, France, 2010, 19-26.

[5] R. Parasuraman and C. D. Wickens, Humans: still vital after all these years of automation. *Human Factors* **50**(3) (2008), 511-520.

[6] *Procedures for Air Navigation Services - Air Traffic Management.* Fifteenth ed. International Civil Aviation Organization, ICAO, 2007.

[7] *Aeronautical Charts. Annex 4 to the Convention on International Civil Aviation.* 11th ed. International Civil Aviation Organization, ICAO, 2009.

[8] *Procedures for Air Navigation Services – Aircraft Operations,* vol. 1 and 2. International Civil Aviation Organization, ICAO, 2006.

[9] *Manual on Volcanic Ash, Radioactive Material and Toxic Chemical Clouds.* Second ed. International Civil Aviation Organization, ICAO, 2007.

[10] J. K. Kuchar and L. C. Yang, A review of conflict detection and resolution modeling methods, *IEEE Transactions on Intelligent Transportation Systems* **1**(4) (2000), 179-189.

[11] W. B. Wong, S. Rozzi, A. Boccalatte, S. Gaukrodger, P. Amaldi, B. Fields, M. Loomes, and P. Martin, 3D-in-2D displays for ATC. In: *6th EUROCONTROL Innovative Research Workshop*, Bretigny, France, 2007, 47-62.

[12] *VatIta Dispatching*, IAC No. 350, Instrument Approach Chart, 2003.

[13] J. C. Mankins, *Technology Readiness Levels*, NASA, Office of Space Access and Technology, Advanced Concepts Office, April 1995.

Probabilistic Networks Basis Criteria of Quality Assurance

Anton BYKAU

Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus

Abstract. The paper considers the application interface testing technology on the basis of probabilistic networks; it offers the criteria of reliability dependence measure for the different application modules sharing the same functionality; it proposes an expand criterion of the testing completeness by the measure of uncertainty of the modules reliability; it is shown how to use automated mechanism of test coverage estimation by using graphical elements general templates; it offers the technology of atomic test union into common model; it proposes a bug reporting mechanism, which allows to estimate the quality of application under test and the quality of automated testing itself.

Keywords. Automated testing, probabilistic networks, criteria of quality assurance

Introduction

The concept of software quality assurance is usually described as a system of attributes or characteristics that can be evaluated using appropriate metrics [1]. These metrics help evaluate the presence of the corresponding characteristic or attribute of quality in software. The main criteria for the quality of the software are listed in ISO 9126.

In this paper, two factors are considered. According to ISO 9126 standard they are: functionality as the ability of software to solve demanded by a user problem, under certain conditions, and reliability as the ability of software to support specific performance under specified conditions [2]. These factors are directly related to software testing.

According to Myers' a test is an execution of a program in order to find the differences between the current state of software and the required (as the evidenced of the error presence) [3]. Also, testing is a process of software evaluation, the assessment of its qualities according to some metrics [4]. The source of information for a testing process could be a source code, a structure of input data, a set of requirements and the model of a developed application. The terms "white box testing" and "black box testing" refer to whether the test is performed with a source code or with interface, both user interface and application programming interface provided by a module under test.

For each test strategy the appropriate criteria for the test coverage could be provided. Test coverage criterion is a metric for completeness of a testing assessment. It measures the variety of situations classes that are taken for testing. The greater the level of test coverage, the greater situations classes are covered (greater verity of situations are taken for a test), the more bugs can be detected.

The most common criteria for a black box are:

- The testing of main application functions.
- Required specification testing
- The testing of classes of input data
- The testing of classes of output data

For the last two criteria there are the following aspects of testing based on the nature of the data used for testing:

- The testing of a tolerance range
- The testing of length of the data set
- The testing of the order of the dataset.

The criteria for white-box testing are based on the source code availability of control flow directly during the execution of the code. The testing is performed according to the logic of the program and the reasonableness of its execution on the basis of the white box criteria. These criteria are:

- Function coverage
- Statement coverage
- Branch coverage
- Decision coverage
- Modified condition coverage

1. Quality assurance criteria modifications

In this paper, it is proposed to use some innovations in the process of automating the functional testing of software through interface. The test automation is performed using both a tester action recorder system and an interface-analyzing algorithm. It is proposed to analyze the composition of GUI elements for each state of the application under test. The analysis is performed using general UI elements patterns. It allows controlling the test coverage of all graphical elements and functions of the program.

It is proposed to use not only black-box testing criteria e.g. functions testing, input and output data classes testing but also white-box testing criteria. The interface state model allows testing in similar way as the white-box testing criteria e.g. combinatorial conditions coverage in interface terms.

It is proposed to classify the graphic elements of modules that share common functionality in order to use common tests. It allows associating the measure of reliability of the modules that share common functionality. As an outcome the assessment of testing completeness could be produced. This mechanism allows evaluating the priorities of running different automated tests. It gives the opportunity to adjust the testing plan (as a milestone of problem solving) by estimating the modules that either are not yet tested either. In other words, the priority of running tests could be estimated.

2. Some problems of testing Web applications

The process of large software systems development usually faces requirements change. As a result, developers are adjusting up to 80% (an average) of application code [5], and automated tests ought to be updated too. The process of maintaining automated tests up to date is one of the most time-consuming tasks of automation [6]. This happens because the automated tests are not integrated or the test code combined incorrectly. Due to the fact that most of application code is combined in libraries, force test developers also combine automated test code into common library. This library is developed specifically for a current project. Each of such libraries has a similar structure, however, the usage of one library on different projects is impossible.

There are some differences between the automation GUI testing process and the application developing. An automated tests developer has an object for tests writing - the application under test. The automated tests developers have to adjust their tests to new application interface versions. This is the cause of a future work planning problem and a bad design of automated tests common functions. The writing of a convenient library is a very complex task. It requires a good further development plans description. Unfortunately, such information is often misses. Also, as a result of a large number of separated updates to the general methods, the common functions code quality becomes worse and worse.

Another problem of GUI test automation is UI elements test coverage estimation. The GUI elements of program under test can be removed and added during a development process. Unfortunately, there are no tools for monitoring such sort of changes. Also, there are no tools for estimation of Web UI tests coverage.

Most projects use Unit testing as a means of GUI test automation, however, for such kind of testing it is necessary to perform a lot of preparation steps. It makes a developer to design dependent tests or included preparation steps before each test. As a result, some steps are performed repeatedly a grate amount times and the result of their work is scattered in various test results.

Another disadvantage of Unit testing is bugs reproduce relation, but it's difficult to relate bug tracking system with an automated GUI test tool. Most of the Unit test frameworks don't allow binding specific drops of tests with a corresponding defect. Falling tests have to be manually excluded from the testing process due to bug fixing.

1. Probabilistic networks basis technology of Web application interface test automation

The using of Selenium test tools for GUI test automation showed that the most important part of the automated test system is a flexible language for defining graphical interface elements. This problem is well solved by using the XPath language. It allows creating flexible templates to find the graphic elements in the DOM of web interface.

A developer should abstract from a concrete UI element on the page and use generic UI element templates and common tests to simplify the automated tests maintenance. UI templates creation is a complex process, but it greatly simplifies the further support of automated test. Also, GUI tests generalization requires test data integration. The generalization of the test data allows applying the same data set for each class of the GUI elements according to their environment.

The generalization of tests using for different modules of application leads to the generalization of test results. The process of the generalization of testing reflects that the same methods of calculation and the processing of the input data are used for different parts of the interface. It looks like the same functionality testing in different environments. If the system receives different test results using the same input data during the testing of various parts of interface that use the same module for data processing, in the results may be a probability describing correctness of module work for all parts of interface. Such probability can be used to predict the same tests fails for other parts of application.

Generalization testing also allows accelerating and reducing the monotony of the automation process. A tester may use a visual code generator classifying UI elements and application state. He could apply the existing element classes of new elements, the same adapting the common tests for the specific elements of the program. A tester could use ready-made data sets to perform a testing process according to the criterion of combinatorial conditions testing.

The code generator algorithm can be represented as a phased process and consists of both the interface analyze phase and the phase of a tester choice a test available for current application state. After the page is loaded test system applies the list of patterns according to the hierarchy of possible nesting graphic elements and creates an interface map. The result is the loading into the hierarchically nested array of rectangular areas and their classes. Finally, the test system creates a context menu designed to select ready-made tests for each UI element class. As a result, there should be enough for a tester to wait the end of the interface analysis procedure, call the context menu of the desired element and evaluate the correctness of the testing. This action menu can be changed by changing appropriate tests and data for UI elements class.

Unit testing automation is useful for white-box testing, but Unit test using as a means of GUI test automation is difficult for maintains. This problem can be solved by using application interface state model. The nodes of such kind of model is the key state of application interface and communications are the set of available operations. Interface model using allows organizing the testing of any available sequence of operations. The model acts as an interpreted description and the kernel of test system appears as interpreter.

Also, the model using allows verifying the testing process in response to the obtained results. The core of the testing system can find alternative ways of preparation steps, in the case of bug in a way of navigation, using a heuristic algorithm to search for alternative paths in the graph [7]. Also it allows combine tests, reducing the total time of testing. It leads to the goals based on the automated test system. The system could automatically find and execute the necessary preparation steps according to any test aim. This leads to a server base test system, which can provide the intermediate bug reports and results of testing, continue further testing in parallel. Such system allows updating the application interface model and testing new items, according to a new description without stopping the testing other modules.

Also application interface state model using can solve the bug tracking and automated test system integration problem. The number of states for defect life cycle should be increased up to four. It allows separating the automated tests correctness estimation from software quality estimation. This problem is well described by J. Myers [8].

If a tester found a previously unknown defect, he adds the defect into bug tracking system, adds a negative test or associate defect with the existing test. If related bug is opened and the test fails, the system shows the result in green for tester’s team member and in red for application developers. This means that automated test working is proper and the application under test contains a bug. This way, tester implements self-testing functionality [9].

After the bug has been fixed the automated test system shows successfully passed test in green for application developers and in red for tester's team member. This behavior means that the defect is no longer reproduced, and its status should be changed to "closed". If the defect appears again the test system will automatically determine the cause of fails.

The result of the technology of UI test automation is an interface states model. This model consists of atomic tests union so that the end state of the previous testing phase becomes the initial state for the next testing phase. If a test is used for multiple functions testing or for the different application states, this common test corresponds to the several appropriate connections or the model. This generalization of the testing reflects the code reusing during application development. For these tests the results of testing could be expanded by the measure of correctness of module work for all parts of interface. Also, each link of application state a model (test) related to measure the uncertainty of the testing results for the current version of application. Such probability can be considered as the conditional probability. The condition appears both already obtained results of other similar tests, and the current test result for the previous version, which was obtained by regression testing.

To formalize these kinds of relationships and the probability is proposed to use a probabilistic network. These networks are used as a tool for decision-making, in case of contradictory data. The probabilistic networks is a base technology represented in this paper use a modification algorithm of the R. D. Schechter [10] algorithm. The modified algorithm allows processing the network information even in the presence of features in the network like:

- The communications of probabilistic networks can be translated in a several directions.
- The communication of probabilistic networks can conflict with each other.

The result inside the node can take multiple values. The condition of the nodes is described by the probabilities that the result takes the appropriate value. The sum of all probabilities inside the node is equals to one. It indicates that the response can reliably takes only one value. The fact that the network connection may conflict with each other leads to considering mutual influence of the facts and produces an approximation of the solution.

To describe the algorithm give an example of the algorithm work for two simple networks. For simplicity the results, let use the relationship between only two nodes, and let conditional probabilities equal either 1 or 0. The probability of the target node can be calculated by using the Bayesian formula according to the conditional probability stored in the properly connection

$$P(A) = P(A|B) \times P(B) \tag{1}$$

However, how algorithm calculates probability of the target node if the communication conflict with each other? Let we consider the following example (Figure 1).

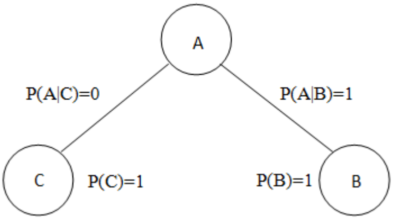


Figure 1. The contradictory evidence

In this example, the C-A and B-A communication can be considered as independent evidences, and the probability of A calculates as the probability of two independent events

$$P(A) = \frac{P(A|B) \times P(B) + P(A|C) \times P(C)}{2} \tag{2}$$

This formula can be applied to any number of connections converging into a single node. Another difficulty is the cycles in the network. Let's change the previously structure of the network connecting C and B. Assume that a given vertex A.

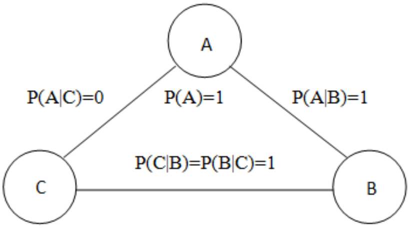


Figure 2. The contradictory relationships

For the network shown in Figure 2 is obvious the contradiction is the link from node A dictate the nodes B and C of the different states, but the connection C - B requires the identity of the values P(C) and P(B) of the nodes.

The decision of the controversy should be removing of one of connections out of the network, but before we do not know the relation to remove, we cannot do this. A temporary solution with equal confidence in the relations should be equal deviation of the node values B and C of dictated by the well-known characteristic P (A) and conditional probabilities - P (C = 1) = P (B = 0) = 0,333, P (C = 0) = P (B = 1) = 0.667.

This solution was obtained by excluding alternately connections out of the network, and averaging the results. Eliminate the link A - B: P (C) = 0, P (B) = 0, eliminate the link A - C: P (C) = 1, P (B) = 1, eliminate the link C - B: P (C) = 0, P (B) = 1.

The advantage of the algorithm is that the relationship can link more than two characteristics, and the algorithm inside the connection is limited by the requirement that for the same initial data algorithm should return the same result. Communication may be a function of several arguments in any programming language. The bidirectional communication between the two characteristics is described by the couple of connections oriented in opposite directions.

2. The main elements of the proposed Web applications testing technology

The technology allows organically combining manual testing and automation. The technology result is an integrated technology model. The model timely reflects the current charges of application under development.

The test system can automatically control the UI elements tests coverage and monitor all interface charges. The test system allows reusing of common tests for a several parts of the application interface with uses common functionality. The Interface states model using allows testing application interface according the criterion of combinatorial conditions testing of white box testing.

The use of probabilistic networks allows estimating the priority of the modules testing, using a measure of the uncertainty the testing results of individual modules, and the results for testing previous version of application.

The paper proposes to use more than two classical states of the life cycle of bug. The testing results are shown depending on the role of the person on the project. Although this innovation is aimed to integrate automated test and bug tracking systems, it allows evaluating the correctness of testing model and quality assurance the application under test separately. Also, this separation allows combining both the positive tests that verify the correct application behavior, and negative tests that verify the absence of incorrect behavior.

3. Criteria for application testing and test coverage estimation

In this paper it is proposed to use generalized patterns for each class of graphical components of Web application interface test automation. The use of generalized pattern of UI elements allows controlling the interface elements test coverage for all application states. It allows test application interface according to black-box criterion of functions testing.

The union of specific interface elements and their groups in the abstraction is performed by a tester as the reflection of functions and modules reuse for web application development. This criterion of test coverage is similar to the conditions combinatorial criterion for testing the white-box testing.

In this paper it is proposed to unite automated atomic tests into common model of application interface. The using application state model allows estimating the program functions reliability by testing application directly, and during the execution of preparatory steps.

The paper discusses the using conditional probability for estimation the connectivity of modules with the using common functionality. This measure can be used for calculation of uncertainty the results individual tests and it combinations according to obtaining the same tests result for other modules. The estimation of uncertainty tests results and tests combinations could be used for increase or decrease the test coverage for associated modules. Such measure could minimize the time of testing overwhelming majority of the functions application under test. It allows quickly verifying reproducing the maximum number of defects, minimizing the testing time. The test system is able to determine the priority of running the tests without human intervention, based on previously obtained results.

4. The main results

The paper discusses the probabilistic networks based on technology using application interface model. The mechanism UI elements test coverage of all application stats is described. The technology uses generalized patterns of UI elements to provide the testing according to black-box criterion of functions testing.

The paper describes the technology of atomic GUI tests combining into the common network. The nodes of the model are all the possible states of the application under test, and arches - all available operations on the interface. The using model of application interface states provides the testing according to criteria similar to the criterion of combinatorial coating conditions for white-box testing.

This technology is tested on the real world Ajax interface application testing, and has proven its effectiveness and convenience in comparison with the unit tests interface automation frameworks.

In this paper, a new measure for estimation the connectivity of test result of the different modules with using common functional is described. The Metrics of modules connectivity was suggested by Larry Constantine [4], which was also an early adherent of such concepts.

In this paper, the measure of priority of the testing the application modules for various combinations of conditions is proposed. The measure allows varying the test coverage of modules, obtained by analyzing the test results for the current or previous versions of the application.

The paper describes the mechanism of bug life cycle tracking, developed and tested by the author, allows evaluating the correctness of the automated tests, and the quality assessment of the application under test separately from each other. This separation is necessary because the system is automated testing and may contain bugs or may not correspond to the changed application requirements like any other program.

In this paper, a procedure that allows you to prepare the data to testing of tolerance range, testing of classes of input and output data is described.

Acknowledgement

The author thanks his scientific adviser I. Piletski for his help in preparing this paper.

References

[1] D. M. Marks, *Testing Very Big Systems*, Bellcore (McGraw-Hill), New-York, 1992.

[2] *ISO/IEC 9126-1, Software engineering – Product quality – Part 1: Quality model*, Geneva, Switzerland, ISO, 2001.

[3] G. J. Myers, *The Art of Software Testing*, Finance and Statistics, Moscow, 1982.

[4] V. V. Kulemin, *Methods of Software Verification*, Institute for System Programming, Moscow, 1995.

[5] I. Vinnichenko, *Automation of Testing Processes*, Peter Press, C-Petersburg, 2005.

[6] *Automated Testing Info* [Internet]. Available from: http://automated-testing.info/content/samoe-slaboe-zveno-vashey-avtomatizacii.

[7] S. Russell, P. Norvig, *Artificial Intelligence: a Modern Approach*, Williams, Moscow, 2007.

[8] G. J. Myers, *The Art of Software Testing*, John Wiley & Sons, Inc., New Jersey, 2004.

[9] J. Bicevskis, Z. Bicevska, and J. Borzovs, Regression testing of software system specifications and computer programs. In: *Proceedings of the 8th Software Quality Week*, San Francisco, paper 5-T-1, 1995.

[10] R. D. Shachter, Evaluating influence diagrams, *Operations Research* **34**(6) (1986), 871–882.

Design and Implementation of the Heterogeneous Multikernel Operating System

Yauhen KLIMIANKOU

Department of Computer Systems and Networks, Belarusian State University of Informatics and Radioelectronics, Belarus

Abstract. The design of the computer system was significantly changed due to the emergence and popularization of the multicore processors. Moving to the advanced multicore processors, moving to the heterogeneous computer systems and increasing of the integrity level between computer system components are the main trends of the computer systems development. Significant changes in the computer systems design make reasonable the attempt of reviewing the operating system design to make it optimal for the new hardware platform. The proposed operation system design assume moving from monolithic centralized operating system to the decentralized network of the distributed independent nodes, each of which will play the role of the processor driver and threads container. The proposed design provide the numbers of the benefits against ordinal operating systems: dynamics in space and time, improved level of reliability and flexibility, support of the heterogeneous computer systems.

Keywords. Multiprocessor computer system, heterogeneous computer system, real-time system

Introduction

Computer systems design is changing much faster than the operating systems design. The internal architecture of the modern computer resembles a distributed network system consisting from the mix of processor cores, caches, internal communications, I/O devices and expansion cards. The modern computer is similar to the early parallel computer systems or multiprocessor systems of the last century. Multi-core computer systems, which are essentially the same multi-processor systems localized on the processor die, occupy an increasingly strong position in most segments of the computer market. Switching to multi-core computer systems is one of the main trends of the computer development, because processor manufacturers have already come close to the limits of the transistor microelectronics and further performance gains in this direction is no longer possible. Further growth of processor performance can be achieved only by intensive arrangements in core design development, and extensive arrangements in increasing number of processor cores. Research and development in the field of multi-core processors were considered as the main direction of processors development by leaders of the processor market, Intel, AMD and ARM. The same way follow the other companies producing processors such as Oracle (CPU SPARC) and IBM (CPU POWER, Cell etc.). Moreover, the development of multi-core processors has the following trends: multiple increase in the number of processor cores located on the same die, transition from UMA-architecture to the NUMA-architecture, using more productive and advanced local inter-core interfaces and buses, and system interfaces and buses, the differentiation of processor cores and transition to the heterogeneous design of the computer system, equipping of main and peripheral components of computer system by advanced specialized processor cores. Thus, all goes towards decentralization and differentiation. It should also be noted that transition to an advanced multi-core is accompanied by a significant increase of the system bus traffic, so we should expect that in future the central system bus will be replaced by a distributed network of buses with NUMA memory architecture. The significant change in the computer system hardware platform leads to the significant change in the operating system design [1, 2]. Moving to the multicore processor based computer system can be considered as the next such significant change. Moreover, the Andrew Tanenbaum notes that the research of the multi-processor focused operating system design is one of the most promising [1].

Utilizing computing power of multicore processors is impossible without the support from the operating system side. Operating system designed for single-processor computer systems on multiprocessor computer systems will use only one processor of the entire array of those available. However, modern operating systems, written with support of multiprocessor computer systems, only mask the problem by involving other processors in the computational process, but trying to do these as if still working in a uniprocessor environment, in such a way that only few kernel modules are designed with the true multiprocessor support in the mind. This leads to the inefficient processor cashes utilization and extensive locking of the system bus. The technology of modern computers and processing units, as well as trends and prospects, assumes development and use of new technologies of computing. And apparently, these technologies will be a symbiosis of computing technology for central processing units, massively parallel processors, parallel and distributed computing and networking technologies.

Basic element of the software is the operating system. It is a complex software system connecting the rest of the software with the hardware platform and represents the software system architecture designed to deal with user tasks and ensuring the autonomy of the computer system. Thus, the main functions of the operating system include the following:

- implementation of organized interaction between tasks and hardware;
- equitable sharing of hardware resources (especially RAM and CPU time, as well as access to services, hardware etc.) of computer system among all tasks;
- providing reliable and stable computer system autonomy.

We can safely say that the technology of operating systems did not keep pace with changes in its hardware platforms - processors. If we carefully look at the IT-industry, it becomes clear that the developers of operating systems are kind of conservative on the IT market. This is probably partly due to the constant need to maintain backwards compatibility for a lot of software written for ordinary operating system. But on the other side gap between the manufacturability of hardware and software platforms can not grow indefinitely. Gradual mass migration to multicore processor device is a new opportunity to review the operating system architecture in which considering lessons learned in the theory of operating systems, and the accumulated innovations in processor architecture and the tendency of their further development are needed. The most progressive in this regard is the concept of heterogeneous multi-core operating system.

1. Next generation Operating System Design

The main idea of the proposed operating system design is moving from the centralized monolithic operating system design to the design of the decentralized distributed network of operating system nodes with high level of autonomy.

The monolithic kernel can be divided to the numbers of separated true microkernels [3] each of which runs on its processor core. The system memory also divided between all operating system kernels in that way that each kernel control and manages its own part of the system memory and don't interfere with other kernels. As result one of the main system abstractions is kernel that incorporates the computation and memory resources with the policies of its management, and from one point of view represents one of the independent operating system node communicating with the other system nodes, and from another point of view plays role of the platform and container from the separated tasks. Migration of the task between operating system nodes during balancing of performance consuming processes is an explicit process managed by external policies that locates outside of the kernels. By default all tasks (threads) related to the same context (process) executes on the same kernel.

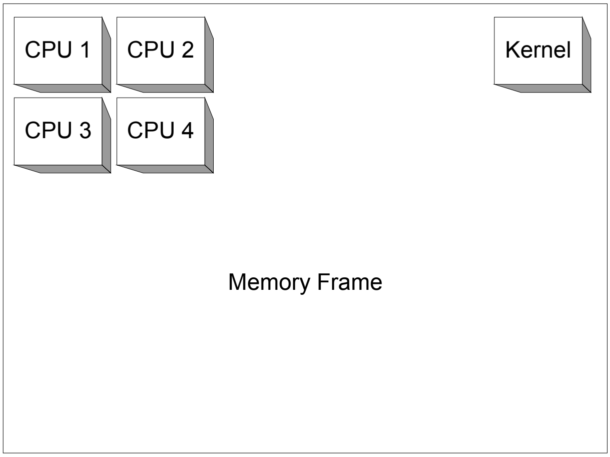


Figure 1. Hardware resources from the point of view of the commodity operating systems

Traditional operating systems design looks at the major hardware resources (CPU and memory) in the way that shown on the Figure 1. The single operating system kernel manages all system resources. From the point of view of the heterogeneous multikernel operating system design the computer system hardware is a set of the simple single processor quasi computer systems each of which includes its own CPU and its own amount of the memory (Figure 2). Consequently operating system is a network of the independent nodes each of which includes the already mentioned above simple quasi computer system and operating system kernel that manages it.

Independence of the operating system nodes is one of the major attributes of the proposed operating system design, because it provides not only distributed management of the system resources, but also divides all operating system to the set of the security domains, that in its turn will lead to the improved flexibility, security and reliability of the overall system.

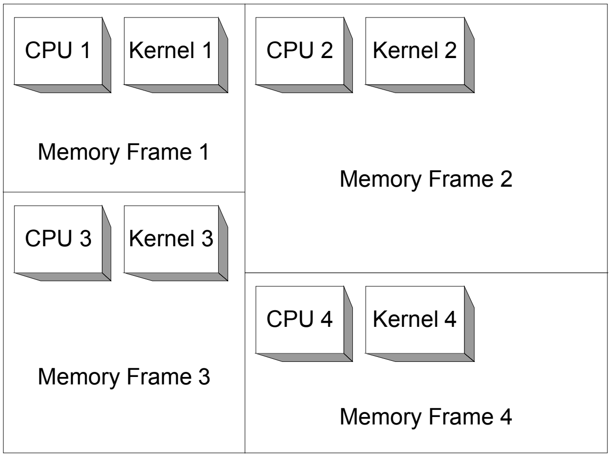


Figure 2. Hardware resources from the point of view of the heterogeneous multikernel operating system

Moving from the single kernel operating system design to the multikernel operating system design require including of the new function of the limited set of the functions of the classical microkernel's – inter-kernel communication. Implementation of this function on the kernel side links all separated kernels into a single operating system network. From the point of view the network consist from the number of communication scopes, one of which is global. Each of the operating system nodes by default is included to the global scope and to the arbitrary number of the other communication scopes. Upon attaching to the scope kernel automatically get access to the broadcast communication channel between all kernels of this scope and can register unicast communication channel. Implementation of the communication channels is based on the message passing on the shared memory region enforced by inter-processor interruptions.

It must be noted that the presented operating system design can be used for building operating systems of four types:

- Homogenous operating system on the homogenous hardware platform. (The classical SMP system where all kernels deliver the same interface.)
- Homogenous operating system on the heterogeneous hardware platform. (The system from mix of the different processors, where all kernels deliver the same interface.)
- Heterogeneous operating system on the homogenous hardware platform. (The classical SMP system, where API provided by kernels can differ from kernel to kernel.)
- Heterogeneous operating system on the heterogeneous hardware platform. (The system from mix of the different processors, where API provided by kernels can differ from kernel to kernel.)

Actually building OS from the source code could be replaced with constructing OS from the set of already compiled kernel binaries, which can be delivered by different distributors.

From the other point of view dividing operating system to the set of independent nodes improves the flexibility, security and reliability. Because the compromising of one of the kernels in the system from the one hand is almost impossible due to microkernel design and from another hand don’t compromise the overall system. In the same way, crash in the kernel from one hand is almost impossible due microkernel design and from another hand the kernel crash won’t stop all the system.

2. Conclusion

This paper presents the research work focused on the field of operating system and kernel design. The purpose of the work is to develop the operating system design and architectural approaches for the most actual computer systems: heterogeneous computer systems based on multi-core processors and embedded, real-time computer systems focused on fully automatic, robotic and cyber-physical systems. The focus of our research is to get the qualitative changes benefits based on the hardware platform improvements in contrast to the goals of the Barellfish project focused on the performance and scalability [4]. The raw micro-kernel has been developed to check the proposed operating system design and other kernel features. Unfortunately, the limited amount of time and lack of hardware prevented the full performance analysis and comparison with other modern kernels.

The main result of this research project is concept of heterogeneous multi-kernel operating system design. This concept propose not only minimization of the kernel and excluding as much functionality as possible from it, but also splitting the traditionally sole kernel to the set of significantly independent micro-kernels, each utilizing one of the processors or processor cores in the system. In addition it proposes switching from centralized operating system kernel to decentralized set of the operating system kernels. As a result, it suggests to look at the operating system as at the decentralized distributed computer tightly coupled network. Also it tries to look at the parallelism problem from another point of view, at which application executes on the same processor all the time by default, and can be extended or moved to another processor only by direct request to the native kernel and approving of request from the kernel on with the application want to come. This approach provides the numbers of advantages:

- The kernel become simpler due to multiprocessor synchronization becomes unnecessary.
- The most of applications running on the computer systems don't need so big computation power to use more than one processor.
- Locking application on one processor offers more productive utilizing of processor cache and relieves locking and traffic of the CPU-RAM bus.
- Proposed design makes it easy to port operating system from the most popular SMP computer systems to the more perspective NUMA computer systems.
- Provides great level of configurability due to proposed design base on the microkernel architecture concept and inherits all its key advantages.

Along with already mentioned advantages, the proposed design of the heterogeneous multi-kernel provides a number of the unique features, that can't be produced by the modern operating systems.

The first advantage is native portability and supporting of heterogeneous computer systems that seems to be a very perspective way of computer systems development. We can just mention such projects as IBM Cell processor and project of hybrid notebook that incorporates both x86 and ARM processors at the same time. Development of processors with specialized cores looks as very attractive idea along with the achievements in transistor microelectronics limits. One more point must be noted according to this advantage. The rapid development of FPGA technologies makes it possible to create computer system that will combine the traditional processors with FPGA chips. Such computer systems will have possibility to form optimized specialized for some computation processor cores on the fly. Such systems would require rapid changing of the operating system kernel and other software that will be executed on the soft processor at the FPGA chip.

The second key advantage provided by proposed heterogeneous operating system design if full software dynamics. It means that in such systems all software components can be replaced on the fly without any rebooting or shutdown of the computer system and without stopping the service of clients. At the one hand, in long term perspective, it allows making full software upgrade, moving to the new operating system, patching the kernel and any other such stuff on the fly. At the other hand, in short term perspective, change the operating system features, properties and behavior on the fly in response to the rapidly changing requirements. For example, it will be possible to switch from general purpose or media focused operating system kernel to the kernel that can satisfy real-time requirements and after some critical work switch back to the general purpose kernel. It also will be possible to switch to the real-time mode not all computer system but only some part of the system. In perspective, it can lead to creation such type of the computer system in which it will be possible to change, add or remove any hardware and software components including processors and memory without shutdown (something like advanced plug and play technology that will cover all computer system but not only several types of the peripheral devices).

And finally the third key advantage provided by proposed heterogeneous operating system design is unprecedented fault tolerance based on combination of key microkernel design advantages, excellent fault tolerance features that are consequence from the decentralized autonomous nature of the design and advanced opportunities for critical functionality duplicating. It will be possible to build computer system based on this design that can be broken only due to overall hardware fault (for example, power off). The possible software faults on all levels, starting from level of applications and ending on the kernel level can be recovered without data loss and only with small loss of time estimated by several microseconds.

References

[1] A. S. Tanenbaum, *Modern Operating Systems*, 3rd edition, Prentice Hall, 2007.

[2] В. Г. Олифер, *Сетевые операционные системы*, Питер, 2009.

[3] J. Liedtke, Improving IPC by kernel design. In: *Proceedings of the 14th ACM Symposium on Operating System Principles (SOSP)*, Asheville, NC, December 1993.

[4] A. Baumann, The multikernel: a new OS architecture for scalable multicore systems. In: *Proceedings of the 22nd ACM Symposium on OS Principles*, Big Sky, MT, USA, October 2009.

SIP Protocol as a Communication Bus to Control Embedded Devices

Ramunas DZINDZALIETA
Institute of Mathematics and Informatics
Akademijos str. 4, Vilnius Lithuania
ramunas.dzindzalieta@gmail.com

Abstract. SIP (The Session Initiation Protocol) is widely used as the signaling protocol for various services in the omnipresent environment. SIP is textual based protocol so the ability to process SIP messages quickly is critical for the performance of consumer electronic devices, such as SIP phone and the various Gateway. Our work enables homogeneous communication between heterogeneous distributed devices. Using SIP protocol communication bus to control widely varying entities, including ZigBEE, serial based (RS232) X10 and other based devices.

Keywords. SIP, communication protocol, embedded system

Introduction

In the network, SIP is implemented in many devices such as the Gateway and various terminals [1]. We propose a ubiquitous service system in network based on SIP, which provides multimedia services and remote control services. It takes use of SIP mobility to preserve service session, even though end-user moves from one computing environment to others. The architecture is implement using SIP and gateway are connected to a sensors environment, such as temperature sensors to realize that at anywhere or SIP phone and home servers connected to Internet. Network device share the ability to interact with physical environment (by sensors or actuators) and have a computing power limited.

The platform of software and hardware is based on the SIP. The main aim is to develop sensor network that will be control, monitored or keep a continuous record of something remotely by smart devices using SIP as a container for collecting data.

1. SIP Project Overview

We want to prove that SIP is good basis to form a communication bus between heterogeneous devices. Extensibility: SIP is request/response protocol, transport-independent and text-based. Like HTTP, SIP is extensible in terms of methods, headers, and message payload. Message payload is format irrespective and SIP supports most kind of data (e.g., SDP, presence information, and SOAP or XML). SIP has provided communication forms, such as commands (RPC-like based on instant messaging), events, and sessions of data streams. SIP platforms are already widely deployed in various forms, because it is standard for IP telephony. SIP protocol become available in various devices such as SIP phone and other including dedicated IP based systems.

2. Building SIP Adapters

We use of SIP as a communication bus between sensors and heterogeneous distributed systems. Let us testing how sensors need to be adapted to connect them to the SIP communication bus.

2.1. Architecture of a SIP Adapter

We must supply entrance to its functionalities via SIP – accepting mechanisms. To entrance to it functionalities are defined of the three interaction modes available in SIP: commands (i.e., control device and status query), events (i.e., event publishing and subscription) and sessions (i. e., invitation to a session of data stream) [2, 3]. Need to request and receive data that may have different formats: command-parameter values (e.g., SOAP), event values (e.g., XML format) and session capability descriptions (e. g., using plain text SDP).

SIP adapters to provide a useful function with an interpreter. The layer obtain from the payload of a SIP message for a given SIP methods, the constituent parts of the corresponding interaction mode (i.e., session, command or event). For example, a SIP request with a MESSAGE method corresponds to a command interaction. The payload interpreter then provides from the request payload a SOAP message, indicating the command name (e.g., *getSensorsParameter*) and the parameter values.

3. Enabling SIP Communication

SIP project consists of two main components:

- a hardware platform made of sensors and actuators connected to Ethernet shield embedded systems with network capabilities;
- specific software on embedded systems implementing the SIP protocol to communicate with the sensors;

The sensor hardware gateway which is based on open source hardware:

- The Atmega board running with I/O connections (USB, serial line) and an Ethernet board to act as a SIP gateway.
- The sensors connected via serial line connection.
- Applications have already been developed to drive sensors on the Atmega board and to collect data from sensor via SIP messages.

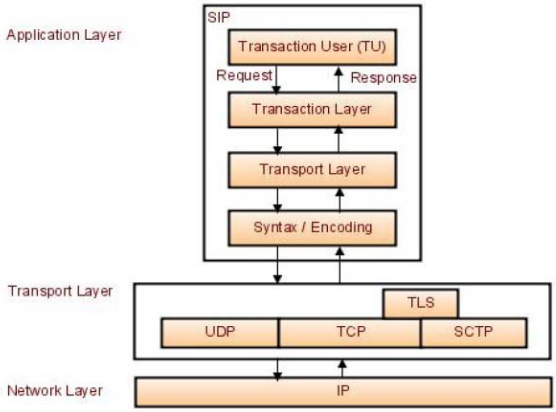


Figure 1. SIP layer protocol

SIP is a layer protocol, comprising the syntax and encoding, transport, transaction, and transaction user (TU) layers [4]. Syntax and encoding layer specifies message format and structure. The transports layer defines how SIP entities send/receive messages over the networks. Above the transport layer is the transaction layer, and the TUs (SIP entities except for the stateless proxies) are in the top layer. SIP transaction layer is the most important layer as it is for request-response matching and retransmission handling.

In SIP, the names found in the fields 'To:' and 'From' are encoded as Universal Resource Locators (URLs). It is useful to define a new type of URL without changing the nature of the protocol to enable better discovery of the device's network address. The structure of the existing SIP address: <entity>@<location> is maintained even when extended to accommodate the devices' names.

A web browser is a software application for retrieving, presenting, and traversing information resources on the World Wide Web. An information resource is identified by a Uniform Resource Identifier (URI) and may be a web page or other piece of content. Hyperlinks present in resources enable users easily to navigate their browsers to related resources. A web browser can also be defined as an application software or program designed to enable users to access, retrieve and view documents and other resources on the Internet.

The board also can connect to a wired network via Ethernet. When connecting to a network, you will need to provide an IP address and a MAC address. The Ethernet Library is fully supported. Using the Ethernet, device will be able to answer a HTTP request. When navigating to your Ethernet shield's IP address, will respond data for a browser to display the input values from all analog pins as Figure 2.

The power pins are as follows:

VIN. The input voltage to the board when it's using an external power source (as opposed to 5 volts from the USB connection or other regulated power source). You can supply voltage through this pin, or, if supplying voltage via the power jack, access it through this pin.

5V. The regulated power supply used to power the microcontroller and other components on the board. This can come either from VIN via an on-board regulator, or be supplied by USB or another regulated 5V supply.

3V. Supply generated by the on-board regulator. Maximum current draw is 50 mA.
GND. Ground pins.
Digital I/O Pins. 14 pins (of which 4 provide PWM output)

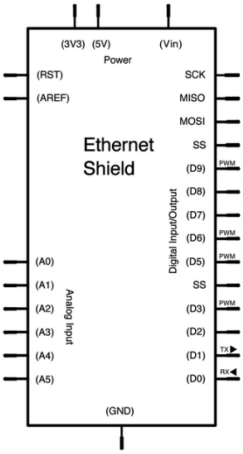
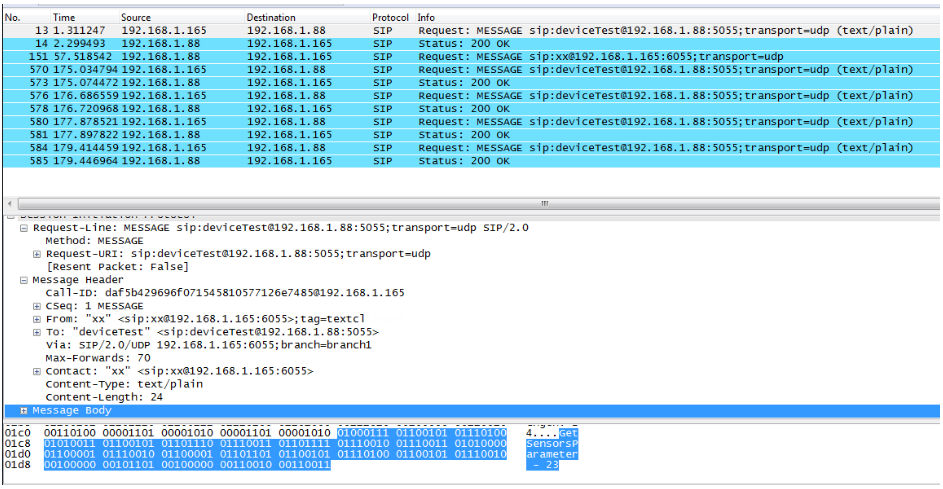


Figure 2. Ethernet shield

The SIP stack that was chosen to develop applications is the oSIP stack to SIP application writing. It is written in C language, is very portable. oSIP supports many transport protocols such as TCP, UDP, and TLS (Transport Layer Security). The GNU oSIP library is written in C and gets no dependencies except the standard C library [5]. oSIP is thread safe and will generally be used in a multi-threaded application. Nevertheless, this is optional. Software was implemented libosip2parser librarary from http://www.gnu.org/s/osip/ has been modified to compile and run on the AtMega board. oSIP is little in size and code and thus could be used to implement IP soft-phone as well as embedded SIP software. oSIP is not limited to endpoint agents, and can also be used to implement "SIP proxy".

oSIP does not intend to provide a high layer API for controlling “SIP Session” at this step [6]. Instead, it currently provides an API for the SIP message parser, SDP message parser, and library to handle “SIP transactions” as defined by the SIP document.



```
No.  Time       Source         Destination    Protocol Info
  13 1.311247   192.168.1.165  192.168.1.88   SIP      Request: MESSAGE sip:deviceTest@192.168.1.88:5055;transport=udp (text/plain)
  14 2.299493   192.168.1.88   192.168.1.165  SIP      Status: 200 OK
 151 57.518542  192.168.1.88   192.168.1.165  SIP      Request: MESSAGE sip:xx@192.168.1.165:6055;transport=udp
 570 175.034794 192.168.1.165  192.168.1.88   SIP      Request: MESSAGE sip:deviceTest@192.168.1.88:5055;transport=udp (text/plain)
 573 175.074472 192.168.1.88   192.168.1.165  SIP      Status: 200 OK
 576 176.686559 192.168.1.165  192.168.1.88   SIP      Request: MESSAGE sip:deviceTest@192.168.1.88:5055;transport=udp (text/plain)
 578 176.720968 192.168.1.88   192.168.1.165  SIP      Status: 200 OK
 580 177.878521 192.168.1.165  192.168.1.88   SIP      Request: MESSAGE sip:deviceTest@192.168.1.88:5055;transport=udp (text/plain)
 581 177.897822 192.168.1.88   192.168.1.165  SIP      Status: 200 OK
 584 179.414459 192.168.1.165  192.168.1.88   SIP      Request: MESSAGE sip:deviceTest@192.168.1.88:5055;transport=udp (text/plain)
 585 179.446964 192.168.1.88   192.168.1.165  SIP      Status: 200 OK

 ⊟ Request-Line: MESSAGE sip:deviceTest@192.168.1.88:5055;transport=udp SIP/2.0
     Method: MESSAGE
   ⊞ Request-URI: sip:deviceTest@192.168.1.88:5055;transport=udp
     [Resent Packet: False]
 ⊟ Message Header
     Call-ID: daf5b429696f071545810577126e7485@192.168.1.165
   ⊞ CSeq: 1 MESSAGE
   ⊞ From: "xx" <sip:xx@192.168.1.165:6055>;tag=textcl
   ⊞ To: "deviceTest" <sip:deviceTest@192.168.1.88:5055>
     Via: SIP/2.0/UDP 192.168.1.165:6055;branch=branch1
     Max-Forwards: 70
   ⊞ Contact: "xx" <sip:xx@192.168.1.165:6055>
     Content-Type: text/plain
     Content-Length: 24
 ⊞ Message Body

01c0  00110100 00001101 00001010 00001101 00001010 01000111 01100101 01110100   4....Get
01c8  01010011 01100101 01101110 01110011 01101111 01110010 01110011 01010000   SensorsP
01d0  01100001 01110010 01100001 01101101 01100101 01110100 01100101 01110010   arameter
01d8  00100000 00101101 00100000 00110010 00110011                               - 23
```

Figure 3. SIP Message between devices

The goal of the project was to make the client (smart phone for examples android) and the servers (our Atmega board) communicate. First of all, the client had to contact the server. Once the communication was established, the server regularly sent the message with temperature data it collected from the sensors. The Android platform can be applied to embedded systems such as control devices with sensors. Mostly embedded systems, C and C++ languages are used as effective languages to control devices, it's very portable and has very low footprint. Embedded devices will be controlled via the Android platform. Developers should create applications by using Java language. But applications written in Java are slower than applications written in native C/C++ languages when the program needs to execute complex operations.

4. Experimental Study

The main aim of this work is to design and implement an automation platform based on SIP. Experiments have been made as in Figure 4. In this environment was introduced by various automation objects, sensors of motion or ranging from smart phone to home devices. This platform is used as a vehicle to experiment with various scenarios. For example, we have developed a ruling application that involves 220V relay, X10 alarms, SIP phones and other sensors as Figure 4. We developed experimental platform which consists: SIP gateway that directly connected with different sensors. The data of sensors is available through the SIP communication bus. For the implementation, we used Atmega board with 16 MHz processor, 4 MB flash memory and Ethernet board.

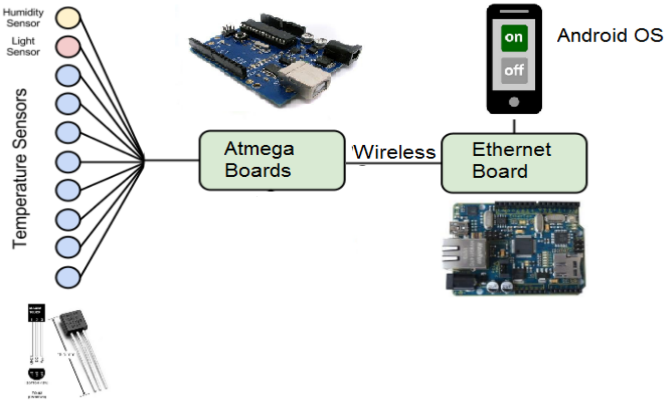


Figure 4. SIP communication bus architecture

5. Conclusions

We presentation homogeneous communications between distributed objects. This presentation relies on the use of SIP as a communication bus for pervasive computing environments. We described a programming support to integrate various objects to the SIP communication. These objects have then been integrated into different of applications for automation systems.

References

[1] B. Bertran, A SIP-based home automation platform: an experimental study. In: *Proceedings of 13th International Conference on Intelligence in Next Generation Networks "Beyond the Bit Pipes"*, IEEE, 2009, 1-6.
[2] B. Campbell, J. Rosenberg, and H. Schulzrinne, C. Huitema, and D. Gurle, *Session Initiation Protocol (SIP) Extension for Instant Messaging*, RFC 3428, The Internet Society, December 2002.
[3] A. B. Roach, *Session Initiation Protocol (SIP) – Specific Event Notification*, RFC 3265, The Internet Society, June 2002.
[4] J. Rosenberg and H. Schulzrinne, *Reliability of Provisional Responses in the Session Initiation Protocol (SIP)*, RFC 3262, The Internet Society, June 2002.
[5] A. Moizard, *The GNU oSIP Library*, 2012. Available from: http://www.gnu.org/software/osip/.
[6] J. Rosenberg, H. Schulzrinne, and G. Camarillo, *SIP: Session Initiation Protocol*, RFC 3261, The Internet Society, June 2002.

Quality of Service: Concept Analysis

Jolanta MILIAUSKAITĖ[1]
Vilnius University, Institute of Mathematics and Informatics

Abstract. This paper summarizes the state of affairs of the doctoral research. Different views of the concept of quality are examined, e.g. product-based, user-based or manufacturing-based view. Finally, the term quality of service is discussed in consideration to web service compositions.

Keywords. Quality, quality of service

Introduction

The main objective of this paper is to summarize the state of affairs of the doctoral research that is in its preliminary phase. The goal of this research is analysis and improvement of web service composition methods. Web service compositions have been investigated by many researchers and a number of different approaches and methods have been proposed in this area. Also a number of open standards facilitating composition of web services have been developed by OMG, OASIS and other standardization bodies. However the achieved results still leave a lot of room for various improvements. In other words, the problem of composition still has not been solved ultimately and more research work, especially on composition of semantic web services is needed. The doctoral research presented concentrates on the methods of quality-driven composition of semantic web services. In this approach quality constraints and preferences are assigned to composite services in addition to the functional requirements. Thus the component services should be selected in such a way that their composition would meet those constraints and preferences in the possible best way. It means that it is possible only to approximate the required quality and that a number of acceptable solutions exist. So, the best solution should be chosen. Besides, this problem is domain-dependent one because the quality of service (QoS) in each application domain is described in terms of different quality characteristics. Up to date, general web service quality model still is under development. In addition, even in cases when only domain-independent characteristics of quality are taken into account, not all quality requirements and preferences can be decomposed, allocated to component services and flowdowned to component service. It is unknown which domain independent quality characteristics can be expressed by characteristics of candidate component services. One more problem is that web services run in highly dynamic environment, in which the number of component services providing required

[1] Jolanta Miliauskaitė: Doctoral Student, Akademijos 4, LT-08663, Vilnius, Lithuania; E-mail: jolanta.miliauskaite@mii.vu.lt

functionality and acceptable QoS is constantly changing. Consequently the run time selections of component service from candidates choose at design time should be done.

In this phase of doctoral research, we aims to define in more precise way the term *Quality of Service,* to understand why the current methods that are used to predict or/and evaluate the QoS for composition of semantic web services are still not enough efficient and to state the aims and tasks of further research. In order to achieve this goal, the conceptual analysis of the concepts *Quality* and *QoS* should be done. We attempt to answer the question: In which extent it is possible to generalize the different meanings of the term *Quality of Service* and to build the ontological model that defines this concept in a formal way? According to Guarino [1], concepts "*have an internal structure, as they "bundle" together further concepts or binary relations (roles)*". By ontological model of QoS we mean formal representation of internal structure of this concept. The paper surveys briefly the results of the analysis and generalizes the findings.

1. What Is the Quality?

To answer the question, what is the meaning of the term "*Quality of Service",* is far not a simple task. According to Lafuente [2], the concept of Quality of Service still remains flue and its definitions depend strongly on the research context and are not adapted to other contexts. Up to date, there is no common understanding even of the term "*quality*". The five different meanings of the term quality have been distinguished by [3] and investigated by a number of other authors [4, 5, 6, 7, 8, 9, 10]:

Transcendental (or metaphysical) view: According to this view, quality is synonymous with "innate excellence." It is both absolute and universally recognizable, a mark of uncompromising standards and high achievement [3]. However, it means that quality is something that can be recognized but not defined or, in other words, it is simple, unanalyzable property, an ideal, towards which we should strive but which can never be achieved in objective reality. Perhaps the best description of quality based on the transcendental view has been done by Pirsig: in this famous philosophical novel on the nature of quality:

"...even though Quality cannot be defined, you know what it is. ... Quality is neither a part of mind, nor is it a part of matter. It is a third entity which is independent of the two. ... Quality isn't a substance. Neither is it a method. It's outside of both. ... It's the goal toward which method is aimed" [11].

However, such view of quality has little practical utility, may be except for advertising campaigns. Such quality cannot be measured by experts and can be perceived through experience alone. Besides, attributes of "ideal" quality are changing over time. Thus, in real-world situations it is more practical to define quality in some constructive way. Although such definition defines only some approximation of ideal quality, it may be more useful.

Product-Based View: This view defines quality on the basis of quantifiable and measurable characteristics or attributes. In other words, it examines the quality from inside perspective and assumes that a product which has good internal properties has also good external properties. For example, this view is supported in the ISO-9004 standard, where quality is defined as fitness for use, performance, safety and dependability. Sometimes this view is called quality of design [12]. The disadvantages of this approach are that it do not take into account preferences of a particular user and assumes that the absence or presence of an attribute implies higher quality. In other words, it assumes that the greater the amount of a desired attribute possessed by a thing, the higher is the quality. According to Leffler, "*Quality refers to the amounts of the unpriced attributes contained in each unit of the priced attribute*" [13].

User-Based View: This view defines quality as fitness for purpose. Sometimes this view is called market-place quality or consumer preference [12]. It is based on the idea that quality is an individual matter, and things that best satisfy user preferences have the highest quality. Thus this understanding of the quality is context-dependent because the judgment about the quality of a thing depends on the aims and goals for which this thing is intended to be used. Despite the fact that the user-based view is highly subjective, it is more concrete as transcendental view because it is based on such product characteristics as usability, reliability, performance and efficiency evaluated from the user's point of view. So, this view equates customer satisfaction with the quality. According to it, product is of high quality if it satisfies a large number of users.

The user-based view is perhaps the most common view about the quality and even in the definitions of quality given by originators of quality management theory, such as Crosby and Juran the quality is defined as "*conformance to user requirements*" [14] or as "*fitness for use*" [12]. However different users place different weights on the various quality characteristics, So, the most important problems with application of user-based view in practice are that it is unclear how to sum up varying individual preferences of particular users and how to know which attributes are for quality and which for user satisfaction. It is very difficult to develop an unbiased statistical procedure that aggregates such widely varying subjective preferences [15]. In addition, most researchers in the service field claim that, at least for services, quality and customer satisfaction are separate concepts although they share a close relationship [16]. The perceived quality of service tends to be stable construct, whereas a user's satisfaction may change for one transaction to another [15]. User's satisfaction can result from a large number of non-quality issues, such as needs, equity, perceptions of fairness. On the other hand, service quality can be conceptualized as a function of the differences between customer's expectation and performance along the quality dimensions [17]. Thus, user-based definition of quality is the most complex definition of quality.

Manufacturing-Based View: This view defines quality as conformance to requirements specification in which the requirements are stated mostly in technical terms. According to Gilmore, "*Quality is the degree to which a specific product conforms to a design or specification*" [18]. This view is based on the idea that any deviation from the specification decreases quality. Similarly as product-based view, manufacturing-based view defines quality in objective and measurable terms, however, focuses on making error-free products or services but not on the absence or presence of some attributes. Sometimes this view is called quality of conformance [12]. Even though it does not ignore the user's interest in quality, it assumes that this interest can be satisfied if the product is properly constructed. The aim is that a product would be constructed "right the first time" and in such way reworking costs would eliminated or, at least reduced [10]. Manufactured-based view assumes that errors can be eliminated by conformance to process standards and concentrates on engineering and manufacturing practices. It assumes also that product quality can be incrementally improved by improving the process. It is supported by ISO 9001 [19] standard and CMM [20]. However "*process standards guarantee only uniformity of output and may thus institutionalize the production of mediocre or bad products*" [21]. Thus manufacturing-based view concerns about user's needs or preferences only in case when they are correctly identified and reflected in requirements specification. In addition, "*a conformance-to-specifications definition of quality may be inappropriate for services, especially when a high degree of a human contact is involved*" [15].

Value-Based View: This view defines quality as the degree of excellence at an acceptable price. It makes a trade-off between cost and quality, that is, it concerns about providing as much quality as the customer is willing to pay for. The quality thing is one that performs or conforms at an acceptable cost or price [22]. However, according to Garvin,

> *"The difficulty in applying this approach lies in its blending of two related but distinct concepts. Quality, which is measure of excellence, is being equated with value, which is a measure of worth. The result is a hybrid – "affordable excellence"- that lacks well-defined limits and is difficult to apply in practice" [3].*

On the other hand, Boehm argues that

> *"...it is also hard for a value-neutral approach to provide guidance for making its products useful to people, as this involves dealing with different people's utility functions or value propositions. It is also hard to make financially responsible decisions using value-neutral methods" [23].*

Boehm's view is supported also by [10] and many other researchers.

Defining quality as value, it is necessary to consider both the internal conformance to specifications (manufacturing-based view) and the extent to which user expectations are meet (user-based view). A quality model for value based approach has been proposed by Gale [24]. Approaches to customer value measurement have been investigated in [25]. In the context of web services, the Value-Based View can be described by triangle of cost versus functionality versus time to deliver. It means that it is possible to satisfy two of these three factors, but not all three [26].

In summary we can conclude that the concept of quality has multiple and sometimes even muddled definitions, describes a wide variety of phenomena and that meaning of the concept depends on the context and even the time period in which it has been examined. However, new definitions of quality have not replaced old ones. All they continue to be used today. Besides each definition has some strengths and weaknesses and no one is better as others in every situation or context. Consequently, it is impossible to build unique ontological model of the quality in general because we are dealing not with one concept but with the bundle of related concepts. This conclusion is also supported by Reeves and Bednar [15].

2. What Means the Quality of Service?

All early d of quality, up to '70s of twentieth century, services were not explicitly addressed and even later, up to end of twentieth century, most researchers continued to focus on product quality. However in this same time in research works on marketing the attempts have been done to understand and define quality in both manufacturing and service organizations [15]. It was pointed out that, although services and products share many similarities, they differ also in a number of ways: 1) services are intangible, cannot be stocked, and their attributes are difficult to demonstrate (intangibility); 2) services are heterogeneous and it is their fundamental characteristic because results of service varies from day-to-day or from customer-to-customer and of this reason it is hard to standardize their quality (heterogeneity); 3) services are inseparable because to a large extent, they are simultaneously produced and consumed (inseparability); 4) services are extremely perishable, that is, they have zero inventory, cannot be saved for later use, can be used only once else they perish and once sold, they stand sold and cannot be returned (perishability). In addition, a service is a process rather than a thing and consumer's involvement in the production of many services creates additional quality control difficulties for managers. In research literature, above mentioned four services characteristics usually are referred as IHIP characteristics [27]. Although some criticism exists whether services are really different from goods and whether the IHIP characteristics are characterizing services, today, the service concept is operationalized mainly through these characteristics [28]. Even more complicated is the question whether or not existing service concepts, including IHIP characteristics and definitions of quality, are applicable to internet services. Conflicting opinions exist on these issues. For example, Moeller argues that:

> *"The characteristics of intangibility, heterogeneity, inseparability, perishability (IHIP) that have been regularly applied to services have been subjected to substantial criticism, as more and more exceptions occur. The reasons for the criticism are twofold. The focus of services marketing has changed and the development of information and communication technology has advanced dramatically" [27].*

Edvardsson et al. [28] and many other researchers advocate also that technology-based services are, in fact, storable, repeatable, often standardized and last, but not least; the service production does not involve any direct interactions with humans. On the other hand, Hofacker et al. [29] states that e-services are less tangible as traditional services, possible, more heterogeneous, taking into account instability of hardware, software and network environment, highly flexible in terms of physical separation between consumer and producer, and can be stored indefinitely by the provider (on server disk) or user. According to [4], majority of papers written on the topic of internet service quality discuss technical details how to pass information about service quality paying little attention to the meaning of quality itself. However it is obvious that every such context-dependent definition should be compatible with the definition of quality in general as well as with universal software quality model. In [4] authors argue that web service quality model should be based on SQuaRE model for software [30]. They define Quality of Web Service as "*ability of Web Service to provide specific users with specific service in defined context of use*" (user-based view of quality), propose to consider also internal web service quality (manufacturing-based view of quality) and external web service quality (product-based view of quality), and suggest that quality of web service composition should be evaluated taking into account hardware, software environment, service itself and transportation quality characteristics for each candidate service.

3. Conclusions

The analysis of the concept QoS demonstrated that this concept is ambiguous and difficult to define precisely. Although observations that have been made in [4] are very significant and highlight some aspects of the nature of this concept, unfortunately, they are still not enough to develop the ontological model of QoS and more deep analysis of QoS should be done for this aim. It is the aim of our further research.

References

[1] N. Guarino, Formal ontology, conceptual analysis and knowledge representation, *International Journal of Human-Computer Studies* **43**(5-6) (1995), 625-640.

[2] E. L. Hernandez, *Evaluation Framework for Quality of Service in Web Services: Implementation in a Pervasive Environment*, Master Thesis, LIRIS, INSA, Lyon, 2010.

[3] D. Garvin, What does "product quality" really mean? *Sloan Management Review* **26**(1) (1984), 25-43.

[4] W. Abramowicz, R. Hofman, W. Suryn, and D. Zyskowski, SQuaRE based web services quality model. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists* **I**, Hong Kong, March 19-21, 2008, 827-835.

[5] B. Edvardsson, Research and concepts: service quality improvement, *Managing Service Quality* **8**(2) (1998), 142-149.

[6] B. Edvardsson, Service quality: beyond cognitive assessment, *Managing Service Quality* **15**(2) (2005), 127-31.

[7] J. R. Evans and W. M. Lindsay, *Managing for Quality and Performance Excellence*, 7th ed., Thomson, Southwestern, 2008.

[8] J. Nankivel, Quality Perspectives in Project Management, *pmStudent*, 2011 [cited 2012 February 14]. Available from: http://pmstudent.com/quality-perspectives-in-project-management/.

[9] J. E. Ross, *Total Quality Management: Text, Cases, and Readings*, Kogan Page, 1994.

[10] S. Kitchenham and S. L. Pfleeger, Software quality: the elisive target, *IEEE Software* **13**(1) (1996), 12-21.

[11] R. M. Pirsig, *Zen and the Art of Motorcvcle Maintenance*, 2nd ed., Bantan Books, 1974 [cited 2012 February 14]. Available from: http://design.caltech.edu/Misc/pirsig.html.

[12] J. Juran and F. Gryna, *Quality Planning and Analysis*, 2nd ed., McGraw-Hill, 1980.

[13] K. B. Leffler, Ambiguous changes in product quality, *American Economic Review* **72**(5) (1982), 956-967.

[14] P. Crosby, *Quality is free*, McGraw-Hill, 1979.

[15] C. A. Reeves and D. A. Bednar, Defining quality: alternatives and implications, *Academy of Management Review* **19**(3) (1994), 419-445.

[16] S. T. Akinyele, Customer satisfaction and service quality: customer's re-patronage perspectives, *Global Journal of Management and Business Research* **10**(6) (2010), 83-90.

[17] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, Reassessment of expectations as a comparison standard in measuring service quality: implications for further research, *Journal of Marketing* **58**(1) (1994), 111-124.

[18] H. L. Gilmore, Product conformance cost, *Quality Progress* **6**(7) (1974), 16-19.

[19] ISO 9001:2008, *Quality Management Systems – Requirements*, International Organisation for Standardization, Geneva, 2008.

[20] M. C. Paulk, Ch. V. Weber, B. Curtis, and M. B. Chrissis, *Capability Maturity Model for Software*, Version 1.1. Technical Report CMU/SEI-93-TR-024 ESC-TR-93-177, Carnegie Mellon University/ Software Engineering Institute, 1993.

[21] I. Tervonen and P. Kerola, Towards deeper co-understanding of software quality, *IEEE Software* **39**(14-15) (1998), 995-1003.

[22] S. T. Akinyele, The spiritual perspective of quality: a scriptural dimension, *IFE PsychologIA* **16**(2) (2008), 62-77.

[23] B. Boehm, Value-based software engineering: overview and agenda, In S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, P. Grünbacher, editors, *Value-Based Software Engineering*, Springer, 2005.

[24] B. T. Gale, *Managing Customer Value: Creating Quality and Service That Customers Can See*, Free Press, 1994.

[25] B. T. Gale and D. J. Swire, *Value-Based Marketing & Pricing*, Customer Value, Inc. 2006 [cited 2012 February 14]. Available from: http://www.cval.com/pdfs/VBMarketingAndPricing.pdf.

[26] N. Baddoo, *CS2 Software Quality module*, Curse notes, University of Hertfordshire, Faculty of Science, Technology and Creative Arts, 1999 [cited 2012 February 25]. Available from: http://homepages.feis.herts.ac.uk/~2com0047/Lecture2WhatisQuality.PDF.

[27] S. Moeller, Characteristics of services - a new approach uncovers their value, *Journal of Services Marketing* **24**(5) (2010), 359-368.

[28] B. Edvardsson, A. Gustafsson, and I. Roos, Service portraits in service research: a critical review, *International Journal of Service Industry Management* **16**(1) (2005), 107-121.

[29] Ch. F. Hofacker, R. E. Goldsmith, E. Bridges, and E. Swilley, E-Services: A synthesis and research agenda, *Journal of Value Chain Management* **1**(1/2) (2007), 13-44.

[30] ISO/IEC 25000:2005 - JTC1/SC7, *Software Engineering - Software product Quality Requirements and Evaluation (SQuaRE)*, International Standardization Organization, 2005.

Automated eContract Negotiation in Web Service Environment: Electronic Contract Management Aspects

Marius ŠAUČIŪNAS
Institute of Mathematics and Informatics, Vilnius University
Akademijos Str. 4, Vilnius, Lithuania
m.sauciunas@gmail.com

Abstract. The paper addresses the electronic contract management problems in automated eContract negotiation among software agents in the web service environment. From the point of electronic contract management, the aim of negotiation process is to automatically form contractual agreements between different parties, coordinating their behavior and facilitating contract execution. The contracts specify the commitments that the involved parties make to each other and that play the important role in their interactions. Therefore in the contract representation language the commitments must be explicitly represented and specify what should be done, if the legal norms defined by policies are violated, and inform the parties about the behavior they could expect from the others. The paper familiarizes with the details of electronic contract representation problem and with approaches which has been proposed to solve this problem. It presents also a critical analysis of the proposed approaches and summarizes their challenges and drawbacks. The paper analyses also one of more advanced conceptual framework of negotiation process from the electronic contract representation perspective, highlights its drawbacks and proposes how to improve this framework.

Keywords. eContract, contract representation language, deontic logic

Introduction

The subject of this paper is the critical analysis of the electronic contract management process among the software agents in the web service environment. At present time the whole contract lifecycle in eBusiness, including the negotiation, preparation of eContract and its acceptation, predominantly is handled manually. In order to develop an electronic contract, humans should not only write and agree upon it but also to translate manually into some computer-readable internal representation [6]. The negotiation, preparation and usage of eContracts still is a challenge.

The electronic contract representation is especially important in the dynamic environments in which prevail the short time contracts. Such contracts have to be dynamically set to meet end-users and service providers' short period needs. In such circumstances, contracts have an intrinsic dynamic and flexible nature and have to regulate independent behavior of diverse parties. Electronic contract preparation and execution facilitation of is one of central issues in the eContracts area.

The goal of the paper is to discuss state of art of the electronic contract representation problem, to highlight the challenges and the drawbacks of the proposed solutions, and to contrasts the conceptual modeling problems of the electronic contract management aspects with the current negotiation process modeling concepts. The main contribution of the paper is the proposal how to improve one of more advanced negotiation process object-oriented modeling framework [11].

1. Electronic Contract Representation Problem

Electronic contract representation problem arises in the context of eContracting. One of the most important requirements in the semantic Web environments that the eContracts should prepared automatically, evaluated, negotiated and executed without human intervention when electronic contract have an intrinsic dynamic and flexible nature and have to regulate independent behavior of diverse parties. It seems that the representation of electronic contracts is one of central issues in the electronic contract monitoring and evaluation too. As pointed in [12], even the most of research in this area focuses on this problem. The contracts specify the commitments that the involved parties make to each other and that play the important role in their interactions. The contract is the statement of intent that regulates the behavior of involved organizations and individuals. Therefore in the contract representation language the commitments must be explicitly represented and specify what should be done, if the legal norms defined by policies are violated, and inform the parties about the behaviour they could expect from the others. A number of such languages – LRC [3], DocLog [16], SweetDeal [5], CEL [17], BCL [13], eContracts [10], etc. – have been proposed.

2. Electronic Contract Representation Languages

LCR (Logic for Contract Representation) is a language for description interaction in multi-agent systems. This language based on branching-time logic, i.e. the formulae in LCR are interpreted over tree-type branching structures that represent all conceivable ways the system can evolve [3]. The formalism behind the language extends branching-time temporal logic with the deontic relations. In LCR, contract clauses are represented as deontic expressions. The violations and sanctions can also be defined in LCR. However, the main purpose of the language is to formalize the behavior of multi-agent system and to relate this behavior to the global objectives of the system. LCR is not intended to be used in the web service environment.

DocLog [16] is an XML based representation language for contract terms. It is based also on the principles of deontic and action logic. Contractual obligations are treated as norms and represented in a semi-formal way using the extended norm frames [9]. Although the DocLog is intended to be used in the eComerce environment and to support the contract negotiation, the language cannot be used in sophisticated Semantic Web environment because it cannot represent exceptions, temporal and some other important aspects of contracts, and is semi-formal.

Quiet different approach is used in the SweetDeal [1][5] approach. It is a rule-based approach for e-commerce business contracts representation. To describe and communicate contracts, the SweetDeal uses a modular set of logic programming rules about agent contracts with exception handling on top of descriptive logic based ontologies, which describe business processes. Such approach enables software agents to automate the creation, negotiation, execution, evaluation of the contracts and reuse contract description for multiple purposes. The motivation for rule-based approach is that rules as knowledge representation formalism is relatively mature, suitable for prescriptive specification and already long time ago integrated into software engineering mainstream techniques. The advantages of rules for representing executables electronic contracts are that rules are relatively easy to modify dynamically and that they are a high abstraction level formalism, which, at least theoretically, is closer to humans' understandability [14].

The start point of the SweetDeal formalism was pure logic programs. Over the pure logic programs the so-called acyclic (non-recursive) courteous logic programs (CLP) have been defined. CLP is a superclass of ordinary logic programs. It is equipped with classical negation and prioritized conflict handling mechanisms [4]. Classical negation is permitted in rule heads and bodies. The courteous approach is a hybrid approach that integrates the ideas of logic programming and general non-monotonic reasoning. The procedural attachments can be attached to logical conditions in the rule antecedents and consequents. CLP with procedural attachments are called situated CLP (SCLP). SCLP expressively extends declarative ordinary logic programs to include prioritized conflict handling. This enables modularity in specifying and revising rule-sets. In the SweetDeal formalism the rules are represented as XML documents [4]. Such representation enhances human readability, supports inclusion and generation of textual information and facilitates parsing. Later this knowledge representation formalism evolved into RuleML family [5]. RuleML was proposed as an alternative to SWRL standard [7]. In summary, a SCLP is suitable to represent fully-specified executable contracts as well as partially-specified contracts that are in the midst of being negotiated [14]. The partially-specified contracts can be viewed as contract templates [14]. The set of negotiables and the structure of a contract in terms of services and attributes are specified by process, contract and other ontologies. Syntactically, the names of predicates appearing in the rules may denote classes and properties in OWL ontology and the names of individuals appearing in rules may refer to individuals in an appropriate ontology. Semantically, the referenced ontological knowledge base is viewed as a background theory for the rule base. Inter allies, ontological knowledge is used for exception handling. During the execution of the contract exception condition (e.g. late delivery, non-payment) could occur, for handling these exceptions process knowledge is required. Thus, ontologies enable to specify more complex contracts with behavioral provisions [5].

The SweetDeal approach is supported by an integrated set of tools, SweetRules [15], that supports creation, evaluation, negotiation, execution and monitoring of formal e- contracts. It provides also a communication protocol between the contracting agents, contract knowledge bases and agent communication knowledge bases. Contract negotiation messages exchanged between the parties are considered as contract knowledge bases that are executable in the SweetRules environment. Each knowledge base consists of six parts: rules, facts, ontologies, effectors, fact-queries, their answers, and conditional queries [1]. Rules describe if-then implications of contractual fragments. Facts are rules without bodies. Ontologies define vocabularies over which the rules are defined. OWL ontologies and rule-based object-oriented default inheritance ontologies are allowed. Effectors are procedural attachments of SLCP. They can execute real-world business process (e.g. e-mail messaging) that are associated with the execution of the contract. Each agent (i.e. negotiating party) has internal knowledge base containing rules that facilitate its communication [1].

In summary, the SweetDeal approach is an well-theoretically-grounded approach that supports many aspects of e-contracting and negotiation. However it does not provide any means to describe deontic modalities and, consequently, is not sufficient to define all legal aspects of negotiated contracts.

Due to volume limitations of this paper other languages will not be discussed in detail.

3. Conceptual Modeling of Negotiation Process and Electronic Contract Management Aspects

3.1. Analysis of the Lin's Conceptual Framework from the Electronic Contract Representation Perspective

The Lin's conceptual framework [11] is one of wide-accepted conceptual models of the negotiation process for Web services contracting. He sees this process as a collaboration of the three conceptual entities: the service requester, the service provider and the service discovery agency.

In the semantic Web environments the eContracts should prepared automatically, evaluated, negotiated and executed without human intervention by software agents. In order to negotiate about contractual agreements, the conceptual model should provide mechanisms to specify contract structure and content, related to contract representation, normative statements, related to involved parties behavior regulation and semantic meaning, related to meaning of contract concepts provision. Lin's conceptual framework [11] does not provide any details how to do this. Most problematic issues are the way in which the framework models the contract manager, and the proposed protocol rules, for the signing of the contract under designated contract template. The model assumes that the contract manager, component of service requester, is maintaining contract template for making agreements. In eContracting environment, where contractual agreement has an intrinsic dynamic and flexible nature, they should be managed and negotiated by several different parties with different features and characteristics. Contracts can never be static, rigid and agreed always under the same contract template unless all template terms which are required for all situations will be defined, but this will be quite difficult to manage for Semantic Web. Another way of contract template using is to combine this solution to some other which enables to form the architecture for automatic contract negotiation. Negotiation mechanisms which have to be used for negotiable parameters of the contract template and to get the final contract in not described either. Contract template structure and content not defined too. Besides, the assumption that the service requester must maintain contract template is questionable. These are obvious drawbacks from the contract management perspective. Another drawback is that parties of automated contract negotiation process have no common understanding on the concepts they agree, i.e. proposed model not deal with semantic meaning on the party's used terminology. Besides supporting of contracts semantics is significant on achieving domain independency in semantic web environments. Another drawback is that the model does not provide any business process monitoring solution. One more assumption in the model that the agency collects the evaluations of service providers' presented by the requesters (trust values in terms of the author) and this information should be enough to evaluate the service, nevertheless for sophisticated evaluation according to agreed contract details this information is insufficient.

3.2. Propositions how to Improve Lin's Conceptual Framework

To adapt the Lin's conceptual framework to the needs of electronic contract management issues, it necessary, first of all, to remove the above discussed drawbacks.

Contract structure and content. The model do not provide any information regarding contract structure and content, consequently some XML-based languages, designed to express contractual agreements in a form, understandable for human beings, could be used for this aim.

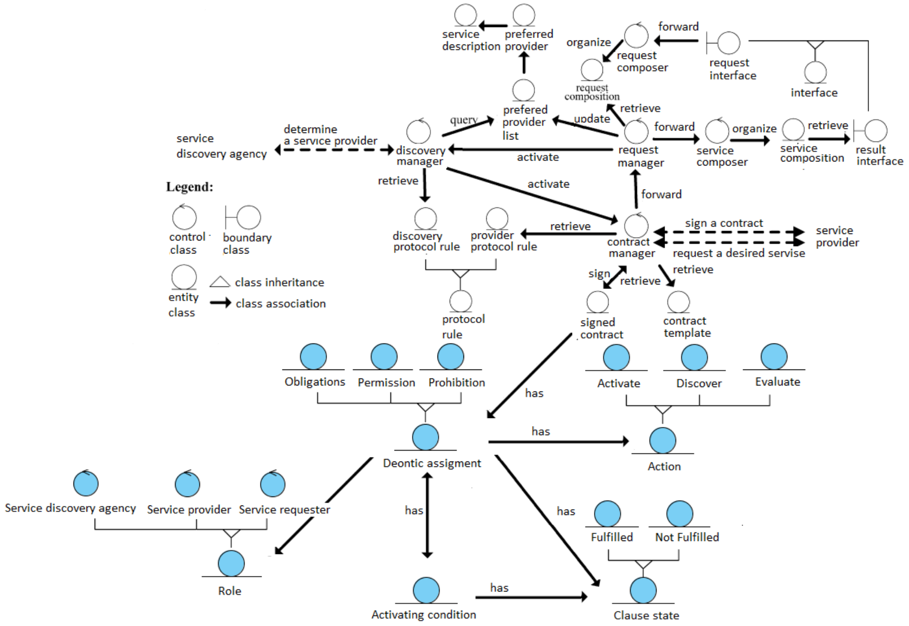


Figure 1. Object class diagram extension for the service requester

Involved parties' behavior regulation and semantic meaning. One of the most important requirements in the context of eContracting is that in semantic Web environments the eContracts should prepared automatically, without human intervention. Contracts should be prepared and established by software agents. Every contract can be modeled as set of different roles, that allocates the tasks to the agents and set of different clauses that regulates the behavior of them. Every agent, depending on the role it is playing in electronic contract, is able or must to perform certain action. The behavior of contractual agents needs to be regulated after contract establishment. For this purpose contract norms, regulating the behavior, can be specified in electronic contract. These normative statements can be modeled based on deontic logic, the logic of the normative concepts, which represents agent's relationship with the concepts of obligation – agent have to do and action, permission – agent is allowed to do a action, and prohibition – agent isn't allowed to do an action. These concepts could be extended by sanction concept - applied in case certain obligation hasn't been fulfilled. In most cases these concepts of deontic logic could be used to model electronic contract. Another serious requirement for eContracts that involved parties should have common terminology and interpretation of the contract concepts they agreed on. To achieve it, ontology, which provides common interpretation in the domain, could be used.

Proposed model extension presents how to incorporate common and domain ontologies. Common ontology provides meaning for general terms, needed for every contract, while domain ontology provide domain meaning to the same terms, described in common ontology, but make them domain independent. Common ontology, as mentioned before, describes the general terms of the contracts e.g. deontic assignments, modeled in deontic logic, specifies roles to perform an certain action. Activating condition specifies conditions, which activates deontic assignment, then current state of the statement could be tracked. Domain ontology extends common ontology terms, e.g. Role class can be specified by three subclasses, dominating in Lin's model, the same rule is valid and for Action class. All these propositions modeled and provided as extension to object class diagram (Figure 1.).

Contract monitoring. The volume of this paper does not allow me to discuss the required solutions in detail. The main idea is that mechanisms similar to that are provided by CONTRACT [8], TPaML [2] or ECL projects and can be used for this aim.

4. Summary and Conclusions

In this paper, the critical analysis of the electronic contract management process among software agents in the web service environment has been performed. In such environment, electronic contracts, with respect to intrinsic dynamic and flexible nature, have to be prepared automatically, evaluated, negotiated and executed without human intervention. The contract is the statement of intent that regulates the behavior of involved organizations and individuals. From the electronic contract management perspective, several significant aspects, such as contract structure, mechanisms to govern collaboration between parties and contract semantic, for specifying complex contracts with behavioral provisions, have to be taken into account dealing with the eContract management problem. From this perspective, major groups of approaches and mechanisms facilitating the electronic contract management problem can be identified: an XML-based languages, languages based on branching-time, deontic, action, programming rules logic, policy-based approaches, role-based approaches. The drawbacks and challenges of each group have been discussed in the paper. Further, the object-oriented Lin's negotiation model [11] that is accepted by many researchers working in the automated negotiation field has been evaluated from the electronic contract management perspective. Its shortcomings have been highlighted, and some significant improvements of the model have been proposed.

The critical analysis of the automated eContract management problem demonstrates, that a lot of different approaches and useful ideas have been proposed up to date, some of them could be distinguished for further investigation and model improvement, e.g. XML-based languages, designed to express contractual agreements in a form, understandable for human beings, role based architectures, contract templates, ontologies, another solutions, not mentioned in this paper due to volume limitation, but which are important to monitoring and evaluation areas. A lot of experimental research should be done for this aim. It intends to be a major focus of my further research.

References

[1] S. Bhansali and B. N. Grosof. Extending the SweetDeal approach for e-procurement using SweetRules and RuleML. In: A. Adi, S. Stoutenburg, S. Tabet, editors, *Proceedings of the First International Conference on Rules and Rule Markup Languages for the Semantic Web (RuleML 2005)*, Galway, Ireland, November 10-12, 2005. Lecture Notes in Computer Science **3791**(2005), Springer, Berlin, 113-129.

[2] A. Dan, D. M. Dias, R. Kearney, T. C. Lau, T. N. Nguyen, F. N. Parr, M. W. Sachs, and H. H. Shaikh. Business-to-business integration with TpaML and a business-to-business protocol framework. *IBM Systems Journal* **40**(1) 2001, 68-90.

[3] V. Dignum, J-J. Meyer, F. Dignum, and H. Weigand. Formal specification of interaction in agent societies. In: M. Hinchey, J. Rash, W. Truszkowski, C. Rouff, and D. Gordon-Spears, editors, *Formal Approaches to Agent-Based Systems (FAABS)*. Lecture Notes in Computer Science **2699** (2003), Springer, Berlin, 37-52.

[4] B. N. Grosof, *Courteous Logic Programs: Prioritized Conflict Handling for Rules.* IBM Research Report RC 20836, December 30 1997, revised from May 8 1997. Available from: http://www.mit.edu/~bgrosof/paps/rc20836.pdf.

[5] B. Grosof, A roadmap for rules and RuleML in the Semantic Web. *IEEE Intelligent Systems* **18**(5) (2003), 76-83. Available from: http://www.mit.edu/~bgrosof/paps/ruleml-ieee-intell-sys-2003.pdf.

[6] P. Hasselmeyer, Ch. Qu, L. Schubert, B. Koller, and Ph. Wieder, Towards autonomous brokered SLA negotiation. In: P. Cunningham, M. Cunningham, editors, *Exploiting the Knowledge Economy: Issues, Applications, Case Studies (eChallenges 2006)*, Barcelona, Spain, October 2006. IOS Press, 44-51. Available from: http://www.hasselmeyer.eu/pdf/echal06.pdf.

[7] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, *SWRL: a Semantic Web Rule Language Combining OWL and RuleML*, 2004. Available from: http://www.w3.org/Submission/SWRL/.

[8] *Contract based Electronic Business Systems State of the Art.* IST Contract Project Deliverable, 2007.

[9] R. van Kralingen, A conceptual frame-based ontology for the law. In: *Proceedings of the First International Workshop on Legal Ontologies*, Melbourne, Australia, 1997, 6-17.

[10] L. Leff and P. Meyer, editors, *eContracts. Version 1.0.* Committee specification. OASIS, 27 April 2007. Available from: http://docs.oasis-open.org/legalxml-econtracts/legalxml-econtracts-specification-1.0.html.

[11] J. Lin, A conceptual model for negotiating in service-oriented environments, *Information Processing Letters* **108**(4) 2008, 192–203.

[12] D. Mobach, *Agent-Based Mediated Service Negotiation*, PhD thesis, Computer Science Department, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, 2007.

[13] S. Neal, J. Cole, P. F. Linington, Z. Milosevic, S. Gibson, and S. Kulkarni, Identifying requirements for business contract language: a monitoring perspective. In: *Proceedings of the Seventh IEEE International Enterprise Distributed Object Computing Conference (EDOC 2003)*, Brisbane, Australia, 16-19 September 2003. IEEE Computer Society, 2003, 50-61.

[14] D. M. Reeves, B. N. Grosof, M. P. Wellman, and H. Y. Chan, *Towards a Declarative Language for Negotiating Executable Contracts*. IBM Research Report, Computer Science, RC 21476 (96914), 11 May 1999. Available from: https://www.aaai.org/Papers/Workshops/1999/WS-99-01/WS99-01-007.pdf.

[15] *SweetRules: Tools for Semantic Web Rules and Ontologies, including Translation, Inferencing, Analysis, and Authoring*, SweetRules, 2005. Available from: http://sweetrules.projects.semwebcentral.org/.

[16] Y. Tan and W. Thoen, DocLog: an electronic contract representation language. In: *Proceedings of 11th International Workshop on Database and Expert Systems Applications (DEXA'00),* 6-8 September 2000, Greenwich, London, UK. IEEE Computer Society, 2000, 1069-1073. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=875159.

[17] X. Wang, E. Chen, D. Radbel, T. Horioka, J. Clark, and G. Wiley, The contract expression language – CEL. In: *Proceedings of the IEEE Workshop on Contract Architectures and Languages*, 20-24 September 2004, Monterey, California, USA. Available from: http://www.contentguard.com/drmwhitepapers/The_CEL.pdf.

Author Index

Jaan AIGRO
Arunas ANDZIULIS
Erika ASNINA
Janis BICEVSKIS
Adomas BIRSTUNAS
Anton BYKAU
Jana CERINA-BERZINA
Vytautas ČYRAS
Jānis DĀBOLIŅŠ
Oxana DOLMATOVA
Uldis DONINS
Dale DZEMYDIENE
Ramunas DZINDZALIETA
Peeter ENNET
Tim FREY
Tomas GRIGALIS
Vytautas JANČAUSKAS
Asnate JANSONE
Girts Karnitis
Giedrius KAUKAS
Hannes KINKS
Yauhen KLIMIANKOU
Veit KÖPPEN
Robert KULLAMAA
Olga KURASOVA
Mindaugas KURMIS
Andreas LÜBCKE
Laila NIEDRITE
Aivars NIEDRITIS
Jolanta MILIAUSKAITĖ
Antanas MITASIUNAS
Olegas NIAKŠU
Boris NOVIKOV
Wendy OSBORN
Janis OSIS
Ott Madis OZOLIT
Saulius RAGAISIS
Reinhard RIEDL
Gunter SAAKE
Ain SALULA
Marius ŠAUČIŪNAS
Laura SAVIČIENĖ
Iztok SAVNIK
Martin SCHÄLER
Arunas STOCKUS
Sandra SVANIDZAITĖ
Vija VAGALE
Vladimir VIIES
Anna YARYGINA
Antanas ŽILINSKAS
Julius ŽILINSKAS
Saad ZAAMOUT
Janis ZUTERS