AI*IA 2012

13-15 JUNE 2012
ROME, ITALY

12th AI*IA Symposium on Artificial Intelligence

Doctoral Consortium

SAPIENZA
UNIVERSITÀ DI ROMA

Preface

The AI*IA Doctoral Consortium (DC) was held as part of the 12th Symposium of the Italian Association for Artificial Intelligence (AI*IA). It took place at University of Rome "La Sapienza" on June 15th, 2012. It follows the very positive experience of the first edition of the DC in the 2010 AI*IA Symposium, according to the spirit of the DCs within main international conferences (such as IJCAI and AAAI).

The Consortium was planned to give Ph.D. students conducting research in the field of Artificial Intelligence an opportunity to interact with established researchers and with other students. Although the DC was designed primarily for students currently enrolled in a Ph.D. program, the call for proposals was also open to exceptions (e.g., students currently in a Master's program and interested in doctoral studies). Students at any stage in their doctoral studies were encouraged to apply, since contributions describing both mature and preliminary work were welcomed.

Eleven contributions were received, of which 9 were accepted for presentation. The Consortium consisted in a poster session in the courtyard of the conference venue. Attendance was open to all AI*IA Symposium participants.

We are pleased with the quality of the contributed research summaries and posters. Submitted works covered a broad spectrum of topics within AI, including machine learning, scheduling and optimization, knowledge representation and reasoning, natural language processing. Most contributions highlighted the practical applications of the investigated AI techniques in cutting-edge fields such as embedded systems design, ubiquitous computing and ambient intelligence, question answering, business process management. All students were able to describe their work in a clear and interesting way, as well as to sustain discussion with competence and maturity.

We are particularly grateful to mentors who took part at the DC. We received very positive responses to our invitations to read the students' proposals and discuss them at the venue. Discussions were very open and mentors provided valuable suggestions and ideas to the students, displaying genuine interest for their work.

In conclusion, we are confident that the DC was a good experience for all participants, promoting mutual exchange of ideas and possibly even future research collaborations.

July 2012

Paolo Liberatore, Michele Lombardi, Floriano Scioscia
DC Co-chairs

Table of Contents

Parameter and Structure Learning Algorithms for Statistical Relational Learning

Elena Bellodi
Supervisor: Fabrizio Riguzzi

ENDIF, Università di Ferrara
Via Saragat, 1 – 44122 Ferrara, Italy
elena.bellodi@unife.it

1 Introduction

My research activity focuses on the field of Machine Learning. Two key challenges in most machine learning applications are uncertainty and complexity. The standard framework for handling uncertainty is probability, for complexity is first-order logic. Thus we would like to be able to learn and perform inference in representation languages that combine the two. This is the focus of the field of Statistical Relational Learning.

My research is based on the use of the vast plethora of techniques developed in the field of Logic Programming, in which the distribution semantics [16] is one of the most prominent approaches. This semantics underlies, e.g., Probabilistic Logic Programs,Probabilistic Horn Abduction,PRISM [16], Independent Choice Logic,pD,Logic Programs with Annotated Disjunctions (LPADs) [17], ProbLog [5] and CP-logic. These languages have the same expressive power: there are linear transformations from one to the others. LPADs offer the most general syntax, so my research and experimentations has been focused on this formalism. An LPAD consists of a finite set of disjunctive clauses, where each of the disjuncts in the head of a clause is annotated with the probability of the disjunct to hold, if the body of the clause holds. LPADs are particularly suitable when reasoning about actions and effects where we have causal independence among the possible different outcomes for a given action.

Various works have appeared for solving three types of problems for languages under the distribution semantics:

- inference: computing the probability of a query given the model and some evidence: most algorithms find explanations for queries and compute their probability by building a Binary Decision Diagram (BDD) [5,15,10];
- learning the parameters: for instance, LeProbLog [6] uses gradient descent while LFI-ProbLog [7] uses an Expectation Maximization approach where the expectations are computed directly using BDDs;
- learning the structure: in [13] a theory compression algorithm for ProbLog is presented, in [12] ground LPADs are learned using Bayesian network techniques.

A different approach from distribution semantics is represented by Markov Logic: a Markov Logic Network (MLN) is a first-order knowledge base with a weight attached to each clause, reflecting "how strong" it is. For this language inference, parameter and structure learning algorithms are available as well, see [14] and [11].

By means of the development of systems for solving the above problems, one would like to handle classical machine learning tasks, such as Text Classification, Entity Resolution, Link Prediction, Information Extraction, etc., in real world domains. An example of Text Classification problem is given by the WebKB dataset, containing the text of web pages from 4 universities; each page belongs to one of four classes: course, faculty, research project or student. Given words on the pages, one wish to infer the class. Entity Resolution concerns the problem of information integration from multiple sources, where the same entities can be differently described. An example is represented by the Cora dataset, containing citations of computer science publications, since citations of the same paper often appear differently.

I have considered the problem of learning the parameters and both the parameters and the structure of LPADs by developing three learning systems: section 2 presents a parameter learning algorithm based on Expectation Maximization, section 3 presents two algorithms for structure learning that exploit the first.

2 Learning LPADs Parameters with the EM Algorithm

An LPAD is composed of annotated disjunctive clauses C_i of the form $h_{i1} : \Pi_{i1}; \ldots; h_{in_i} : \Pi_{in_i} : -b_{i1}, \ldots, b_{im_i}$. where $h_{i1}, \ldots h_{in_i}$ are logical atoms, $b_{i1}, \ldots, b_{im_i}$ are logical literals and $\{\Pi_{i1}, \ldots, \Pi_{in_i}\}$ are real numbers in the interval $[0, 1]$ representing probabilities and sum up to 1.

An LPAD rule containing variables represents a number of "simple experiments", one for each ground instantiation of the rule (obtained by replacing variables with constants of the domain), with the disjuncts as the possible outcomes. For each ground instantiation of a rule *only* one pair $(h : \Pi)$ is chosen; in this way a normal non-disjunctive logic program is obtained, called *possible world*. A probability distribution is defined over the space of possible worlds by assuming independence among the selections made for each rule. The probability of a possible world is the product of probabilities of the individual heads chosen in each rule. The probability of a query Q according to an LPAD is given by the sum of the probabilities of the possible worlds where the query is true.

Example 1. The following LPAD T encodes the result of tossing a coin depending on the fact that it is biased or not:

$C_1 = heads(Coin) : 0.5; tails(Coin) : 0.5 : -toss(Coin), \neg biased(Coin).$
$C_2 = heads(Coin) : 0.6; tails(Coin) : 0.4 : -toss(Coin), biased(Coin).$
$C_3 = fair(coin) : 0.9; biased(coin) : 0.1.$
$C_4 = toss(coin) : 1.$

This program models the fact that a fair coin lands on heads or on tails with probability 0.5, while a biased coin with probabilities 0.6 and 0.4 respectively. The third clause says that a certain coin *coin* has a probability of 0.9 of being fair and of 0.1 of being biased, the fourth one that *coin* is certainly tossed.

Each selection of a disjunct in a ground clause of an LPAD can be represented by the equation $X_{ij} = k$, where $k \in \{1, \ldots, n_i\}$ indicates the head chosen, X_{ij} is a multivalued random variable where i and j indicate the clause and the grounding. A function $f(\mathbf{X})$, built on a set of multivalued variables and taking Boolean values, can be represented by a Multivalued Decision Diagram (MDD), a rooted graph that has one level for each variable. Each node has one child for each possible value of the associated variable. The leaves store either 0 or 1, the possible values of $f(\mathbf{X})$. Given values for all the variables $\mathbf{X}$, a MDD can be used for computing the value of $f(\mathbf{X})$ by traversing the graph starting from the root and returning the value associated to the leaf that is reached. An example of such function is: $f(\mathbf{X}) = \{X_{11} = 1 \vee X_{21} = 2 \wedge X_{31} = 1 \vee X_{22} = 3 \wedge X_{31} = 1\}$. A MDD can be used to represent the set of selections over rules, and will have a path to a 1-leaf for each possible world where a query Q is true. It is often unfeasible to find all the worlds where the query is true so inference algorithms find instead *explanations* for the query, i.e. set of selections such that the query is true in all the worlds whose selection is a superset of them. Since MDDs split paths on the basis of the values of a variable, the branches are mutually disjoint so that a dynamic programming algorithm can be applied for computing the probability of a query by a summation. Usually one works on MDDs with a Binary Decision Diagram package, so one has to represent multivalued variables by means of Boolean variables.

The problem I faced is how to efficiently perform "parameter learning", i.e., using training data for learning correct probabilities Π_{ik}. The technique applied exploits the EM (Expectation Maximization) algorithm over BDDs proposed in [9,8] and has been implemented in the system EMBLEM, for "EM over BDDs for probabilistic Logic programs Efficient Mining" [1,2,3].

EMBLEM takes as input a set of interpretations (sets of ground facts), each describing a portion of the domain of interest, and a theory (LPAD). The user has to indicate which, among all predicates, are target predicates: the facts for these predicates will form the queries for which a SLD-proof is computed; from these proofs a BDD is built encoding the Boolean formula consisting of the disjunction of the explanations for the query.

Then EMBLEM performs an EM cycle, in which the steps of Expectation and Maximization are repeated until the log-likelihood of the examples reaches a local maximum. Expectations are computed directly over BDDs. EM is necessary to determine the parameters Π_{ik} since the number of times that head h_{ik} is chosen is required. The information about which selection was used is unknown, so the "choice" variables are latent and the number of times is a sufficient statistic.

Decision Diagrams are suitable to efficiently evaluate the expectations since the set of selections used for the derivation of the examples can be represented as the set of paths from the root to the 1-leaf.

3 Learning LPADs Structure

The first system developed for LPADs' structure learning is SLIPCASE, for "Structure LearnIng of ProbabilistiC logic programs with Em over bdds" [4]. It learns a LPAD by starting from an initial theory and by performing a beam search in the space of refinements of the theory. The initial theory is inserted in the beam and, at each step, the theory with the highest log likelihood is removed from the beam and the set of its refinements, allowed by the language bias, is built. The possible refinements are: the addition/removal of a literal from a clause, the addition of a clause with an empty body and the removal of it. For each refinement an estimate of the log likelihood of the data is computed by running a limited number of iterations of EMBLEM. The best theory found so far is updated and each refinement is inserted in order of log likelihood into the beam.

I am now working on SLIPCOVER, an evolution of SLIPCASE which first searches the space of clauses and then the space of theories. SLIPCOVER performs a cycle for each predicate that can appear in the head of clauses, where a beam search in the space of clauses is performed: each clause (built according to a language bias) is tested on the examples for the predicate, its head parameters are learned with EMBLEM and the log likelihood of the data is used as its score. Then the clause is inserted into one of two lists of promising clauses: a list of target clauses, those for predicates we want to predict, and a list of background clauses, those for the other predicates. Then a greedy search in the space of theories is performed, in which each target clause is added to the current theory and the score is computed. If the score is greater than the current best the clause is kept in the theory, otherwise it is discarded. Finally parameter learning with EMBLEM is run on the target theory plus the clauses for background predicates.

The two systems can learn general LPADs including non-ground programs.

4 Experiments

We experimented EMBLEM on the real datasets IMDB, Cora, UW-CSE, WebKB, MovieLens and Mutagenesis and evaluated its performances by means of the Area under the PR curve and under the ROC curve, in comparison with five logic-probabilistic learning systems. It achieves higher areas in all cases except two and uses less memory, allowing it to solve larger problems often in less time.

We tested SLIPCASE on the real datasets HIV, UW-CSE and WebKB, and evaluated its performances - in comparison with [11] and [12] - through the same metrics, obtaining highest area values under both.

We have tested the second structure learning algorithm on HIV, UW-CSE, WebKB, Movielens, Mutagenesis and Hepatitis and evaluated it - in comparison with SLIPCASE, [11] and [12] - through the same metrics. It has overcome them in all cases. In the future we plan to test the systems on other datasets and to experiment with other search strategies.

References

1. Bellodi, E., Riguzzi, F.: EM over binary decision diagrams for probabilistic logic programs. In: Proceedings of the 26th Italian Conference on Computational Logic (CILC2011), Pescara, Italy, 31 August 31-2 September, 2011. pp. 229–243. No. 810 in CEUR Workshop Proceedings, Sun SITE Central Europe, Aachen, Germany (2011), http://www.ing.unife.it/docenti/FabrizioRiguzzi/Papers/BelRig-CILC11.pdf
2. Bellodi, E., Riguzzi, F.: Expectation Maximization over binary decision diagrams for probabilistic logic programs. Intel. Data Anal. 16(6) (2012), to appear
3. Bellodi, E., Riguzzi, F.: Experimentation of an expectation maximization algorithm for probabilistic logic programs. Intelligenza Artificiale (2012), to appear
4. Bellodi, E., Riguzzi, F.: Learning the structure of probabilistic logic programs. In: Inductive Logic Programming 21st International Conference, ILP 2011, London, UK, July 31 - August 3, 2011. Revised Papers. LNCS, Springer (2012), to appear
5. De Raedt, L., Kimmig, A., Toivonen, H.: ProbLog: A probabilistic prolog and its application in link discovery. In: International Joint Conference on Artificial Intelligence. pp. 2462–2467. AAAI Press (2007)
6. Gutmann, B., Kimmig, A., Kersting, K., Raedt, L.D.: Parameter learning in probabilistic databases: A least squares approach. In: European Conference on Machine Learning and Knowledge Discovery in Databases. LNCS, vol. 5211, pp. 473–488. Springer (2008)
7. Gutmann, B., Thon, I., Raedt, L.D.: Learning the parameters of probabilistic logic programs from interpretations. In: European Conference on Machine Learning and Knowledge Discovery in Databases. LNCS, vol. 6911, pp. 581–596. Springer (2011)
8. Inoue, K., Sato, T., Ishihata, M., Kameya, Y., Nabeshima, H.: Evaluating abductive hypotheses using an em algorithm on bdds. In: Boutilier, C. (ed.) Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI). pp. 810–815. Morgan Kaufmann Publishers Inc. (2009)
9. Ishihata, M., Kameya, Y., Sato, T., Minato, S.: Propositionalizing the em algorithm by bdds. In: Late Breaking Papers of the International Conference on Inductive Logic Programming. pp. 44–49 (2008)
10. Kimmig, A., Demoen, B., De Raedt, L., Santos Costa, V., Rocha, R.: On the implementation of the probabilistic logic programming language problog. Theory and Practice of Logic Programming 11(2-3), 235–262 (2011)
11. Kok, S., Domingos, P.: Learning markov logic networks using structural motifs. In: International Conference on Machine Learning. pp. 551–558. Omnipress (2010)
12. Meert, W., Struyf, J., Blockeel, H.: Learning ground CP-Logic theories by leveraging Bayesian network learning techniques. Fundam. Inform. 89(1), 131–160 (2008)
13. Raedt, L.D., Kersting, K., Kimmig, A., Revoredo, K., Toivonen, H.: Compressing probabilistic prolog programs. Mach. Learn. 70(2-3), 151–168 (2008)
14. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62(1-2), 107–136 (2006)
15. Riguzzi, F.: Extended semantics and inference for the Independent Choice Logic. Logic Journal of the IGPL 17(6), 589–629 (2009)
16. Sato, T.: A statistical learning method for logic programs with distribution semantics. In: International Conference on Logic Programming. pp. 715–729. MIT Press (1995)
17. Vennekens, J., Verbaeten, S., Bruynooghe, M.: Logic programs with annotated disjunctions. In: International Conference on Logic Programming. LNCS, vol. 3131, pp. 195–209. Springer (2004)

A Constraint-based Approach to Cyclic Resource-Constrained Scheduling Problem

Alessio Bonfietti
Supervisor: Michela Milano

DEIS, University of Bologna
V.le Risorgimento 2 – 40136 Bologna, Italy
alessio.bonfietti@unibo.it

The cyclic scheduling problem concerns assigning start times to a set of activities, to be indefinitely repeated, subject to precedence and resource constraints. It can be found in many application areas. For instance, it arises in compiler design implementing loops on parallel architecture[5], and on data-flow computations in embedded applications[2]. Moreover, cyclic scheduling can be found in mass production, such as cyclic shop or Hoist scheduling problems[6].

In cyclic scheduling often the notion of optimality is related to the period of the schedule. A minimal period corresponds to the highest number of activities carried out on average over a large time window.

Optimal cyclic schedulers are lately in great demand, as streaming paradigms are gaining momentum across a wide spectrum of computing platforms, ranging from multi-media encoding and decoding in mobile and consumer devices, to advanced packet processing in network appliances, to high-quality rendering in game consoles. In stream computing, an application can be abstracted as a set of tasks that have to be performed on incoming items (frames) of a data stream. A typical example is video decoding, where a compressed video stream has to be expanded and rendered. As video compression exploits temporal correlation between successive frames, decoding is not pure process-and-forward and computation on the current frame depends on the previously decoded frame. These dependencies must be taken into account in the scheduling model. In embedded computing contexts, resource constraints (computational units and buffer storage) imposed by the underlying hardware platforms are of great importance. In addition, the computational effort which can be spent to compute an optimal schedule is often limited by cost and time-to-market considerations.

My research focuses on scheduling periodic application. In particular, in first place I have worked together with my research group on a Constraint Programming approach based on modular arithmetic for computing minimum-period resource-constrained cyclic schedules [3]. The solver has several interesting characteristics: it deals effectively with temporal and resource constraints, it computes very high quality solutions in a short time, but it can also be pushed to run complete search. An extensive experimental evaluation on a number of non-trivial synthetic instances and on a set of realistic industrial instances gave promising results compared with a state-of-the art ILP-based (Integer Linear Programming)[1] scheduler and the Swing Modulo Scheduling (SMS)[7] heuristic technique. The main innovation of our approach is that while classical *modular*

approaches fix the modulus and solve the corresponding (non periodic) scheduling problem, in our technique the bounds for the modulus variables are inferred from the activity and iteration variables.

The main drawback of this first approach is the underlying hypothesis that the end times of all activities should be assigned within the modulus. Thanks to this assumption, we can reuse traditional resource constraints and filtering algorithms. However the solution quality can be improved by relaxing this hypothesis.

Therefore, we proposed [4] a Global Cyclic Cumulative Constraint (GCCC) that indeed relaxes this hypothesis. We have to schedule all the start times within the modulus λ, but we have no restriction on end times. The resulting problem is far more complicated, as enlarging the modulus produces a reduction of the modular end time of the activities. Figure 1 explains the concept. Suppose the grey activity requires one unit of a resource of capacity 3. If the modulus value is D, then the activity can be scheduled as usual. If the modulus is reduced to C, the starting time of the activity is the same, while the "modular end time" is c and the resource consumption is 2 between 0 and c. If the modulus is further reduced to B the modular end time increases to b. Finally, if the modulus is reduced to A, the modular end point becomes a and the resource consumption is 3 between 0 and a.

In [4] we show the advantages in terms of solution quality w.r.t. our previous approach that was already outperforming state of the art techniques. The experiments highlight that our approach obtains considerably better results in terms of solution quality for high capacity values. Moreover, the results show that, working with acyclic graphs, the GCCC approach obtains an approximately constant resource idle time.

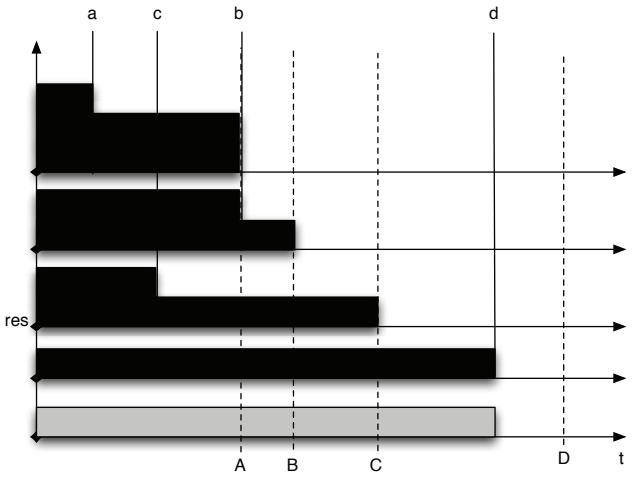

Fig. 1. Resource Profiles w.r.t different modulus values

Further investigation will be devoted to the design of cyclic scheduling heuristic algorithms and their comparison with complete approaches.

References

1. Ayala, M., Artigues, C.: On integer linear programming formulations for the resource-constrained modulo scheduling problem (2010)
2. Bhattacharyya, S.S., Sriram, S.: Embedded Multiprocessors - Scheduling and Synchronization (Signal Processing and Communications) (2nd Edition). CRC Press (2009)
3. Bonfietti, A., Lombardi, M., Benini, L., Milano, M.: A Constraint Based Approach to Cyclic RCPSP. In: CP2011. pp. 130–144 (2011)
4. Bonfietti, A., Lombardi, M., Benini, L., Milano, M.: Global cyclic cumulative constraint. In: Beldiceanu, N., Jussien, N., Pinson, E. (eds.) 9th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CPAIOR'12). Lectures Notes in Computer Science, vol. 7298, pp. 81–96. Springer Verlag, Nantes, France (Jun 2012)
5. Dupont de Dinechin, B.: From Machine Scheduling to VLIW Instruction Scheduling (2004)
6. Hanen, C.: Study of a NP-hard cyclic scheduling problem: The recurrent job-shop. European Journal of Operational Research 72(1), 82–101 (1994), http://www.sciencedirect.com/science/article/B6VCT-48NBGY3-226/2/e869b267f2deeef5b90a18f742a2e2c4
7. Llosa, J., Gonzalez, A., Ayguade, E., Valero, M.: Swing modulo scheduling: A lifetime-sensitive approach. In: pact. pp. 80–87. Published by the IEEE Computer Society (1996), http://en.scientificcommons.org/43187025http://www.computer.org/portal/web/csdl/doi/10.1109/PACT.1996.554030

A Model to Summarize User Action Log Files

Eleonora Gentili
Supervisor: Alfredo Milani

Dipartimento di Matematica e Informatica, Università degli Studi di Perugia
via Vanvitelli 1 – 06123 Perugia, Italy
eleonora.gentili@dmi.unipg.it

Abstract. Social networks, web portals or e-learning platforms produce in general a large amount of data everyday, normally stored in its raw format in log file systems and databases. Such data can be condensed and summarized to improve reporting performance and reduce the system load. This data summarization reduces the amount of space that is required to store software data but produces, as a side effect, a decrease of their informative capability due to an information loss. In this work we study the problem of summarizing data obtained by the log systems of chronology-dependent applications with a lot of users. In particular, we present a method to reduce the data size, collapsing the descriptions of more events in a unique descriptor or in a smaller set of descriptors and pose the optimal summarization problem.

1 Introduction

During last years we have seen an impressive growth and diffusion of applications shared and used by a huge amount of users around the world. Network traffic data log systems, alarms in telecommunication networks and web portals which records the user activities are examples of chronology-dependent applications (CDA) producing in general large amount of data in the form of log sequences.

But log files are usually big and noisy, and the difficulty of finding patterns is very high as well as the number of patterns discovered can be very large [6]. For this reason, data in log files can be condensed and summarized to improve reporting performance and analyze the system behavior.

In this work, a new method to produce a concise summary of sequences of events related to time is presented, which is based on the data size reduction obtained merging time intervals and collapsing the descriptions of more events in a smaller set of descriptors. Moreover, in order to obtain a data representation as compact as possible, an abstraction operation allowing an event generalization process (as in [5]) is defined. The summarized time sequence can substitute the original time sequence in the data analysis.

The reduction of the amount of data produces also as a side effect, a decrease of their informative capability due to an information loss. For this reason, we formally define the summarization problem as an optimization problem that balances between shortness of the summary and accuracy of the data description.

1.1 Related Works

In [5], Pham et al. propose an algorithm that achieves time sequence summarization based on a generalization, grouping and concept formation process, maintaining the overall chronology of events. Our summarization method overcomes the time limitation of this procedure, using time intervals, instead only time instants.

In [6], [4], [3], authors propose methods to produce a comprehensive summary for the input sequence, focused on the frequency changes of event types across adjacent segments. In particular, in [4], Kiernan and Terzi rely to the Maximum Description Length (MDL) principle to produce the summarized time sequence balancing the shortness of the summary and accuracy of the data description. Moreover, in [3], the presented framework summarizes an event sequence using inter-arrival histograms to capture the temporal relationships among events using the MDL. Unlike these works, where are presented methods which partition the time sequence into segments and study global relationships on each segment, and among segments, our method takes into account an overview of the time sequence to solve the summarization problem.

In [1], Chandola et al. formulates the problem of summarization as an optimal summarization problem involving two measures, Compaction Gain and Information Loss, which assess the quality of the summary.

2 The model

In this work we assume that each event is described by a set of users which made actions over objects either at a particular instant or during an interval. Moreover, the assumed time model is discrete, where events are instantaneous and are representable by a tuple (u, a, o, t), with which we describe *users*, *actions*, *objects* and *time* of the actions.

In order to represent more general situations and potentially aggregate information of similar events, we define events as tuples involving sets of users, actions and objects possibly occurred during an interval.

Definition 1. *An event descriptor is a t-uple $X = (U, A, O, I, \delta)$ representing a set of actions A made by a set of users U over a set of objects O, during a given time interval I according to the covering index δ that is defined as the ratio by the number of points in which the actions in A are actually executed and all the points of I.*

Example 1. $X = (\{admin\}, \{login, logout\}, \{IT\}, [10, 50], 0.30)$ represents that *admin* made *login* and *logout* in ITcourse in the 30% of the points of $I = [10, 50]$.

We assume that the labels in the sets U, A and O can be organized in taxonomies, which are organized in hierarchies with multiple inheritance: each taxonomy is associated to an *abstraction operator* ($\uparrow$), allowing to climb the hierarchy.

The abstraction operator applied to a node of the taxonomy returns all the fathers of the node, while, when it is applied to a set of nodes $S = \{s_1, \ldots, s_n\}$, the result is a set $\uparrow(S) = S'$ where at least a s_i is substituted with $\uparrow(s_i)$. Let's two different sets S_1 and S_2, the *minimal abstracting set* S is defined as the first not null set of common ancestor of S_1 and S_2 computed by climbing the taxonomy graph associated to S_1 and S_2, i.e. such that $S = \uparrow(S_1) = \uparrow(S_2)$.

Definition 2. *A time sequence $\boldsymbol{X} = (X_1, \ldots, X_m)$ is a sequence of m event descriptors X_i; m is called size of the time sequence, or data size.*

Given a *time sequence*, we aim to provide methods to reduce its data size, collapsing the descriptions of more events in a smaller set of event descriptors.

Definition 3. *Let's Ω the set of all event descriptors, $X_1 = (U, A, O, I_1, \delta_1)$ and $X_2 = (U, A, O, I_2, \delta_2)$ two event descriptors, $I_1 = [t'_1, t''_1]$ and $I_2 = [t'_2, t''_2]$, we define the merging operator as*

$$\begin{gathered} \oplus : \Omega \times \Omega \to \Omega \\ ((U, A, O, I_1, \delta_1), (U, A, O, I_2, \delta_2)) \mapsto (U, A, O, I, \delta) \end{gathered} \tag{1}$$

such that $I = [min(t'_1, t'_2), max(t''_1, t''_2)]$ and

$$\delta = \frac{\delta_1|I_1| + \delta_2|I_2| - \min(\delta_1|I_1|, \delta_2|I_2|, |I_1 \cap I_2|)}{|I|} \tag{2}$$

The *merging operator* collapses intervals of *event descriptors* with identical label sets. δ is computed considering that events happening in both I_1 and I_2 coincide as much as possible; it is simple to prove that $\delta \leq \max(\delta_1, \delta_2)$.

Example 2. Given the time sequence $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$, where

$X_1 = (\{user_1, user_2\}, \{login\}, \{objA\}, [1, 1], 1.0), X_2 = (\{user_2\}, \{send\}, \{objA\}, [2, 2], 1.0),$
$X_3 = (\{user_1\}, \{read\}, \{objA\}, [4, 4], 1.0), X_4 = (\{user_2\}, \{send\}, \{objA\}, [5, 5], 1.0).$

We can apply the *merging operator* to X_2 and X_4 obtaining

$$X_{24} = X_2 \oplus X_4 = (\{user_2\}, \{send\}, \{objA\}, [2, 4], 0.67).$$

The new *time sequence* $\mathbf{X} = \{X_1, X_{24}, X_3\}$ has a smaller size but less information about events. And no more *merging operation* can be applied.

Definition 4. *Let's Ω the set of all event descriptors and given an event descriptor $X = (U, A, O, I, \delta)$, the abstraction operator $\uparrow_S$ is defined as*

$$\begin{gathered} \uparrow_S : \Omega \to \Omega \\ (U, A, O, I, \delta) \mapsto (U', A', O', I, \delta) \end{gathered} \tag{3}$$

where $S \in \{U, A, O\}$ and $S' = \uparrow(S)$ is obtained applying the abstraction operator.

The *abstraction operator* will make mergeable *event descriptors* having different label sets, generalizing labels in the sets U, A, O.

Definition 5. *Let's $\{X_i : X_i = (U_i, A_i, O_i, I_i, \delta_i), i = 1, \ldots, n\}$ a set of event descriptors, $X_i^* = (U, A, O, I_i, \delta_i)$ is the minimal abstracting event for X_i if each label set U, A, O is the minimal abstracting set respectively for $\{U_i\}$, $\{A_i\}$, $\{O_i\}$.*

For instance, let consider the taxonomy graphs depicted in Fig.1, and given

$$X_1 = (\{user_1\}, \{Create.folder, Save\}, \{log_1\}, I_1, \delta_1),$$
$$X_2 = (\{user_1, user_2\}, \{Disk.op, Write.mail\}, \{log_1\}, I_2, \delta_2),$$

the two *minimal abstracting events* for X_1 and X_2 are $X_1^* = (\{user\}, \{User.op\}, \{log_1\}, I_1, \delta_1)$ and $X_2^* = (\{user\}, \{User.op\}, \{log_1\}, I_2, \delta_2)$ respectively.

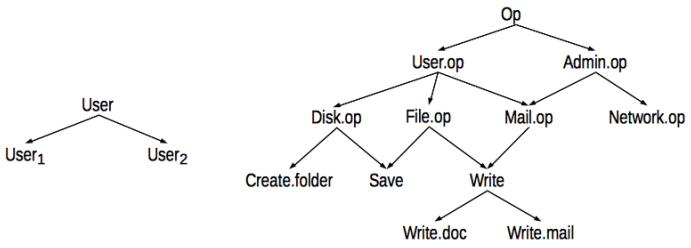

Fig. 1. An example of taxonomy graphs respectively over the sets U and A.

Considering that creating a summary produces information loss, the optimal summarization problem aims to maximize the reduction of data size minimizing the information loss. It is clear that the optimal summarization is a question of tradeoff between application of the *merging operator* and the *abstraction operator* to the *event descriptors*.

3 The optimal summarization problem

Let's $\mathbf{X}_0$ the time sequence of the initial volume of data and $\mathbf{X}$ a summarized time sequence, we define some metrics to assess the quality of $\mathbf{X}$ with respect to $\mathbf{X}_0$.

Definition 6. *The compaction gain of* $\boldsymbol{X}$ *is define as* $\mathcal{C}(\boldsymbol{X}, \boldsymbol{X}_0) = \frac{|\boldsymbol{X}_0|}{|\boldsymbol{X}|}$.

Definition 7. *Given* $I_i = [t'_i, t''_i]$ *and* $G_i = [t''_i, t'_{i+1}]$, *the covering accuracy of* $\boldsymbol{X}$ *is*

$$\mu(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} \delta_i |I_i| + \sum_{i=1}^{m} |G_i|}{\sum_{i=1}^{n} |I_i| + \sum_{i=1}^{m} |G_i|}. \tag{4}$$

The gap intervals G_i are considered as intervals with $\delta_{G_i} = 1.0$, because there are no events happening in each G_i. It easy to prove that $0 \leq \mu(\mathbf{X}) \leq 1$. In particular, $\mu(\mathbf{X}) = 1$ is verified if and only if $\delta_i = 1$, $\forall i = 1, \ldots, n$.

Definition 8. *Given* $\boldsymbol{X}$, *the description accuracy of* $\boldsymbol{X}$ *is defined as*

$$\eta(\boldsymbol{X}) = \min_{X \in \boldsymbol{X}} \left(\min(\omega_U \eta(U), \omega_A \eta(A), \omega_O \eta(O))\right), \tag{5}$$

where $\omega_U, \omega_A, \omega_O \geq 0$ *are the weights of the label sets, and*

$$\eta(S) = \min_{n \in S} \frac{d(r, n)}{h(n)},$$

where $S \in \{U, A, O\}$, r *is the root of the taxonomy* T_S *and* $h(n)$ *is the longest distance from* n *to a leaf.*

Note that $0 \leq \eta(S) \leq 1$. In particular, $\eta(S) = 1$ is verified when n is a leaf, and $\eta(S) = 0$ when n coincides with the root.

Definition 9. *Given $\boldsymbol{X}$ and $\boldsymbol{X}_0$, the information loss of the summarization process is defined as*

$$\mathcal{I}(\boldsymbol{X}, \boldsymbol{X}_0) = \alpha(\mu(\boldsymbol{X}_0) - \mu(\boldsymbol{X})) + \beta(\eta(\boldsymbol{X}_0) - \eta(\boldsymbol{X})). \tag{6}$$

Definition 10. *Given $\boldsymbol{X}_0$ and a real number $\gamma > 0$, we define the Optimal Summarized Time Sequence $\overline{\boldsymbol{X}}$ such that the parameterized ratio between $\mathcal{I}(\boldsymbol{X}, \boldsymbol{X}_0)$ and $\mathcal{C}(\boldsymbol{X}, \boldsymbol{X}_0)$ is minimal, i.e.*

$$\overline{\boldsymbol{X}} = \underset{\boldsymbol{X}}{\operatorname{argmin}} \frac{\mathcal{I}(\boldsymbol{X}, \boldsymbol{X}_0)}{[\mathcal{C}(\boldsymbol{X}, \boldsymbol{X}_0)]^{\gamma}}. \tag{7}$$

4 Conclusions and future works

In this work the problem of summarizing data obtained by the log systems of applications with a lot of users is studied. We have presented a new method to produce a concise summary of sequences of events related to time, based on the data size reduction obtained merging time intervals and collapsing the descriptions of more events in a unique descriptor or in a smaller set of descriptors. Moreover, in order to obtain a data representation as compact as possible, an abstraction operation allowing an event generalization process is defined.

Moreover, we are studying about the formalization and the implementation of an optimal algorithm for the *Optimal Summarization Problem*. The idea is to use suboptimal algorithms to obtain the best summarization algorithm for our method.

References

1. V. Chandola and V. Kumar. *Summarization–compressing data into an informative representation*. Knowledge and Information Systems, 12(3):355–378, 2007.
2. E. Gentili, A. Milani and V.Poggioni. *Data summarization model for user action log files*. In Proceedings of ICCSA 2012, Part III, LNCS 7335, pp. 539–549. Springer, Heidelberg, 2012.
3. Y. Jiang, C.S. Perng, and T. Li. *Natural event summarization*. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pages 765–774. ACM, 2011.
4. J. Kiernan and E. Terzi. *Constructing comprehensive summaries of large event sequences*. In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 417–425. ACM, 2008.
5. Q.K. Pham, G. Raschia, N. Mouaddib, R. Saint-Paul, and B. Benatallah. *Time sequence summarization to scale up chronology-dependent applications*. In Proceeding of the 18^{th} ACM conference on Information and knowledge management, pages 1137–1146, 2009.
6. P. Wang, H. Wang, M. Liu, and W. Wang. *An algorithmic approach to event summarization*. In Proceedings of the 2010 international conference on Management of data, pages 183–194. ACM, 2010.

Knowledge Representation Methods for Smart Devices in Intelligent Buildings

Giuseppe Loseto
Supervisor: Michele Ruta

DEE, Politecnico di Bari
via Re David 200 – I-70125 Bari, Italy
loseto@deemail.poliba.it

Abstract. Home and building automation aims at improving features and capabilities of household systems and appliances. Nevertheless, current solutions poorly support dynamic scenarios and context-awareness. The integration of knowledge representation features and reasoning techniques (originally devised for the Semantic Web) into standard home automation protocols can offer high-level services to users. A semantic-based approach is proposed, able to interface users and devices (whose characteristics are expressed by means of annotated profiles) within a service-oriented home infrastructure.

1 Introduction

Home and Building Automation (HBA) –also known as domotics– is a growing research and industrial field, aimed at coordinating subsystems and appliances in a building to provide increased levels of user comfort and manageability, reduce energy consumption and minimize environmental impact. In latest years, the design of smart HBA environments is attracting efforts from several disciplines, including mobile and pervasive computing, wireless sensor networks, artificial intelligence and agent-based software, coalescing into a research area known as *Ambient Intelligence* (AmI) [12]. A crucial issue for feasible and effective AmI solutions lies in efficient resource/service discovery. Current HBA systems and standard technologies are still based on explicit user commands over static sets of operational scenarios, established during system design and installation. Consequently, they allow a low degree of autonomicity and flexibility. These restrictions can be by-passed through the adaptation and integration of Knowledge Representation (KR) formalisms and techniques originally conceived for the Semantic Web. Ontology languages, based on Description Logics (DLs), can be used to describe the application domain and relationships among resources in a way that can support inference procedures and matchmaking processes, in order to satisfy users' needs and preferences to the best possible extent.

2 State of the Art

Currently, most widespread technological standards for HBA –including KNX (www.knx.org), ZigBee (www.zigbee.org) and LonWorks (developed by Echelon

Corporation, www.echelon.com) – only offer static automation capabilities, consisting of pre-designed application scenarios. They do not allow autonomicity in environmental adaptation given a user profile and dynamic context-awareness.

An early approach towards AmI was proposed in [10]. Intelligent agents were used to automate a service composition task, providing transparency from the user's standpoint. Nevertheless, such an approach was based on service discovery protocols such as UPnP and Bluetooth SDP, presenting a too elementary discovery and supporting only exact match of code-based service attributes. Due to the growing interest in reducing energy consumption, several studies on AmI and multi-agent systems have been proposed for energy management and comfort enhancement [3]. Unfortunately, these solutions either require direct user intervention or only support basic interaction between devices, lacking advanced resource discovery and composition capabilities. The use of knowledge representation can allow to overcome such limitations. Knowledge Bases (KBs) will be exploited to enable a user-device interaction and to interconnect household appliances, using different protocols, in order to share services and data, *e.g.*, related to device energy consumption [4]. In [1] an ontology-based domotic framework with a rule-based reasoning module was introduced to manage and coordinate heterogeneous devices endowed with semantic descriptions. The main weakness of the above works is in the presence of static rule sets and centralized KBs. A really pervasive environment requires a different approach, able to deal with the intrinsically dynamic, decentralized and unpredictable nature of AmI.

3 Proposed approach

Framework. A general-purpose framework for HBA has been proposed, supporting semantic-enhanced characterization of both user requirements and services/resources provided by devices. Following pervasive computing spirit, during ordinary activities the user should be able to simultaneously exploit information and functionalities provided by multiple objects deployed in her surroundings. Each device should autonomously expose its services and should also be able to discover functionalities and request services from other devices.

Technologies and ideas are borrowed from the Semantic Web initiative and adapted to HBA scenarios. Semantic Web languages, such as OWL[1], provide the basic terminological infrastructure for domotic ubiquitous KBs (u-KBs) which enable the needed information interchange. The fully exploitation of semantics in user and device description has several benefits which include: (i) machine-understandable annotated descriptions to improve interoperability; (ii) reasoning on descriptions to characterize environmental conditions (context) and to support advanced services through semantic-based matchmaking.

The reference framework architecture, shown in Figure 1, integrates both semantic-enabled and legacy home devices in a domotic network with an IP backbone. Coordination among user agents and domotic agents (representing

[1] OWL Web Ontology Language, version 2, W3C Recommendation 27 October 2009, http://www.w3.org/TR/owl2-overview/

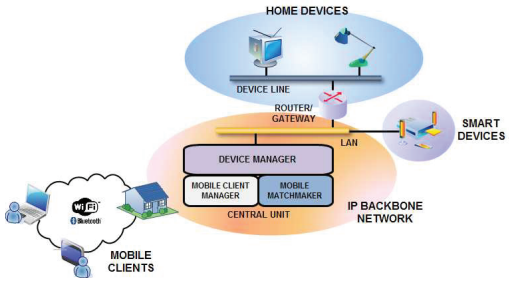

Fig. 1. Framework Architecture

(a) Testbed (b) Device Panel

Fig. 2. Developed Testbed

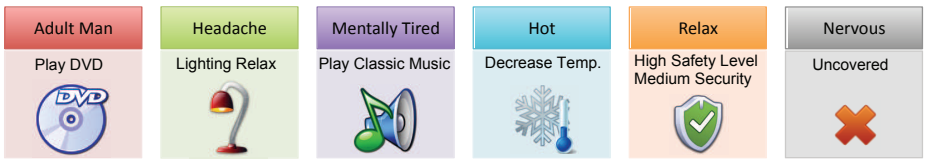

Fig. 3. Concept Covering Result

devices, rooms and areas) is facilitated by a home unit. Communication between client agents and the home system may occur through either IEEE 802.11 or Bluetooth wireless standards. The discovery framework is based on a distributed application-layer protocol. Optimized inference services [2] feature a Decision Support System (DSS) hosted by the coordination unit. Service discovery is not limited to identity matches (infrequent in real situations) but it supports a logic-based ranking of approximated matches allowing to choose resources/services best satisfying a request, also taking user preferences and context into account. Such an approach allows then user to require addressed services instead of simplistic device features. For example, the system could be able to reply to articulated requests as the one reported in what follows: *I am tired and I have a splitting headache. For these reasons, I am very nervous and I wish a relaxing home environment. It is a warm evening and I feel hot.* By sending a request to the home, an accurate home service selection can be performed, as shown in Figure 3. The selected service set includes suggestions for DVD playback and music, following stored user preferences. A lower room temperature and soft lighting settings are selected to improve user comfort. Finally, system sets appropriate home safety and security settings inferred by the mobile matchmaker exploiting the axioms in the ontology. An uncovered part of the request is also present, because there are no specific services able to match nervous user state.

Methodology. In order to grant feasibility, the proposed framework was based on a fully backward-compatible extension of current domotic technologies. Consequently, besides review of the state of the art, the first research phase included a careful study of the most widespread HBA standards, in order to verify the possibility of a semantic enhancement and to select a reference protocol for sub-

sequent work. The second step involved the design of protocol enhancements to support the representation and exchange of semantic information, followed by an extensive evaluation through simulation campaigns. Then the framework has been defined in detail, including: (a) specification of an ontology for the HBA application domain able to support the functional and non–functional requirements of the project; (b) development and optimization of an embedded matchmaking engine, providing standard and non-standard inference services described in [2, 5]. Based on the theoretical framework, a testbed has been developed to evaluate the effectiveness of the approach and to experiment about performance –considering several case studies, with user and device semantic descriptions varying in number and complexity.

4 Results

KNX was selected as reference HBA standard due to its support for multiple communication media, availability of development tools and wide industry acceptance. At protocol level, main contribution includes the definition of new data structures and application-layer services [6] conforming KNX 2.0 specification to store and exchange semantic metadata. Due to the reduced availability of both device storage and protocol bandwidth in current domotic infrastructures, the proposed enhancements envisage the use of a compression algorithm specifically targeted to document in XML-based ontological languages [11]. The mobile semantic matchmaker in [7] has been extended with the Concept Covering inference service [5] –in addition to Concept Abduction and Concept Contraction [2]– to support covering of a complex request through the conjunction of elementary service units. A prototypical testbed, shown in Figure 2, was developed. It represents a small set of home environments equipped with different KNX-compliant off-the-shelf devices. Integration of the semantic-enhanced protocol features in an agent framework is a further step in the research. In [8], main characteristics of the framework are highlighted and early performance evaluation is presented. A subsequent step, under current investigation, involves the exploitation of a semantic-based negotiation protocol seeking to maximize energy efficiency. The agents are able to: (i) negotiate on available home and energy resources through a user-transparent and device-driven interaction; (ii) reveal conflicting information on energy constraints; (iii) support non-expert users in selecting home configurations. The first results are presented in [9].

5 Conclusion and Future Work

A semantic-based pervasive computing approach has been investigated to overcome existing limitations in HBA solutions. The integration of KR and reasoning techniques with current standards and technologies is fundamental to improve user comfort and building efficiency. Enhancements aim at building a distributed knowledge-based framework.

Beyond completing the outlined research tasks, future extensions will include a user agent running on a mobile client, enabling rich and autonomous interactions in a collaborative smart space.

References

1. Bonino, D., Castellina, E., Corno, F.: The DOG gateway: enabling ontology-based intelligent domotic environments. IEEE Transactions on Consumer Electronics 54(4), 1656 –1664 (november 2008)
2. Colucci, S., Di Noia, T., Pinto, A., Ragone, A., Ruta, M., Tinelli, E.: A non-monotonic approach to semantic matchmaking and request refinement in e-marketplaces. International Journal of Electronic Commerce 12(2), 127–154 (2007)
3. Klein, L., young Kwak, J., Kavulya, G., Jazizadeh, F., Becerik-Gerber, B., Varakantham, P., Tambe, M.: Coordinating occupant behavior for building energy and comfort management using multi-agent systems. Automation in Construction 22(0), 525 – 536 (2012)
4. Kofler, M.J., Reinisch, C., Kastner, W.: A semantic representation of energy-related information in future smart homes. Energy and Buildings 47(0), 169 – 179 (2012)
5. Ragone, A., Di Noia, T., Di Sciascio, E., Donini, F.M., Colucci, S., Colasuonno, F.: Fully automated web services discovery and composition through concept covering and concept abduction. International Journal of Web Services Research (JWSR) 4(3), 85–112 (2007)
6. Ruta, M., Scioscia, F., Di Sciascio, E., Loseto, G.: A semantic-based evolution of EIB Konnex protocol standard. In: IEEE International Conference on Mechatronics (ICM 2011). pp. 773 –778 (April 2011)
7. Ruta, M., Scioscia, F., Di Sciascio, E.: Mobile semantic-based matchmaking: a fuzzy dl approach. In: The Semantic Web: Research and Applications. Proceedings of 7th Extended Semantic Web Conference (ESWC 2010). Lecture Notes in Computer Science, vol. 6088, pp. 16–30. Springer (2010)
8. Ruta, M., Scioscia, F., Di Sciascio, E., Loseto, G.: Semantic-based Enhancement of ISO/IEC 14543-3 EIB/KNX Standard for Building Automation. IEEE Transactions on Industrial Informatics 7(4), 731–739 (2011)
9. Ruta, M., Scioscia, F., Loseto, G., Di Sciascio, E.: An Agent Framework for Knowledge-based Homes. In: 3rd International Workshop on Agent Technologies for Energy Systems (ATES 2012) (2012, to appear)
10. Santofimia, M., Moya, F., Villanueva, F., Villa, D., Lopez, J.: Intelligent Agents for Automatic Service Composition in Ambient Intelligence. In: Usmani, Z. (ed.) Web Intelligence and Intelligent Agents, pp. 411–428. InTech (2010)
11. Scioscia, F., Ruta, M.: Building a Semantic Web of Things: issues and perspectives in information compression. In: Semantic Web Information Management (SWIM'09). In Proceedings of the 3rd IEEE International Conference on Semantic Computing (ICSC 2009). pp. 589–594. IEEE Computer Society (2009)
12. Shadbolt, N.: Ambient intelligence. Intelligent Systems, IEEE 18, 2–3 (July 2003)

Molecules of Knowledge: a Novel Perspective over Knowledge Management

Stefano Mariani
Supervisor: Andrea Omicini

Alma Mater Studiorum – Università di Bologna
via Venezia 52 – 47521 Cesena, Italy
s.mariani@unibo.it

Abstract. To face the challenges of knowledge-intensive environments, we investigate a novel self-organising knowledge-oriented (SOKO) model, called *Molecules of Knowledge* (MoK for short). In MoK, knowledge atoms are generated by knowledge sources in shared spaces – compartments –, self-aggregate to shape knowledge molecules, and autonomously move toward knowledge consumers.

1 The *Molecules of Knowledge* Model

The *Molecules of Knowledge* model is a biochemically-inspired coordination model exploiting and promoting self-organisation of knowledge. The basic motivation behind MoK is the idea that knowledge should autonomously aggregate and diffuse to reach knowledge consumers rather than "be searched". The main pillars of MoK are represented by the following (bio)chemical abstractions:

atoms — the smallest unit of knowledge in MoK, an *atom* contains information from a *source*, and belongs to a *compartment* where it "floats"—and where it is subject to its "laws of nature";
molecules — the MoK units for knowledge aggregation, *molecules* bond together somehow-related atoms;
enzymes — emitted by MoK *catalysts*, *enzymes* influence MoK reactions, thus affecting the dynamics of knowledge evolution within MoK compartments to better match the catalyst's needs;
reactions — working at a given *rate*, *reactions* are the biochemical laws regulating the evolution of each MoK compartment, by ruling knowledge aggregation, diffusion, and decay within compartments.

Other relevant MoK abstractions are instead in charge of important aspects like topology, knowledge production and consumption: **compartments** represent the conceptual *loci* for all MoK entities as well as for biochemical processes – like knowledge aggregation and diffusion –, providing MoK with the notions of *locality* and *neighbourhood*; **sources** are the origins of knowledge, which is continuously injected in the form of MoK atoms at a certain *rate* within the compartment sources belong to; **catalysts** stand for knowledge *prosumers*, emitting enzymes which represent their actions which affect knowledge dynamics

within their own compartment, especially to increase the probability of providing him/her with relevant knowledge items.

Atoms Atoms are the most primitive living pieces of knowledge within the model. A MoK atom is produced by a knowledge source, and conveys a piece of information spawning from the source itself. Hence, along with the content they store, atoms should also store some contextual information to refer to the content's origin, and to preserve its original meaning.

As a result, a MoK atom is a triple of the form $atom(\ src,\ val,\ attr\ )_c$ where: *src* identifies unambiguously the source of knowledge; *val* is the actual piece of knowledge carried by the atom – any kind of content –; *attr* is essentially the content's attribute, that is, the additional information that helps understanding it – possibly expressed according to some well-defined ontology or controlled vocabulary –; c is the current *concentration* of the atom, which is essentially the number of atoms of the kind in the compartment.

Molecules Molecules are spontaneous, stochastic, "environment-driven" aggregations of atoms, which in principle are meant to reify some semantic relationship between atoms, thus possibly adding new knowledge to the system—for instance self-aggregated news chunks could shape the conceptual "path" toward a novel interesting news story. Each molecule is simply interpreted as a set of atoms, that is, an unordered collection of somehow semantically-related atoms.

A MoK molecule has then a structure of the form $molecule(\ Atoms\ )_c$ where c is the current concentration of the molecule, and *Atoms* is the collection of all the atoms currently bonded together in the molecule.

Enzymes One of the key features of MoK is that it interprets prosumer's knowledge-related actions as *positive feedbacks* that increase the concentration of related atoms and molecules within the prosumer's compartment, producing the positive feedback required to enable self-organisation of knowledge.

A MoK enzyme has then a structure of the form $enzyme(\ Atoms\ )_c$ where every enzyme – with its own concentration c – explicitly refers to a collection of *Atoms* that the catalyst's actions have in any way pointed out as of interest for the catalyst him/herself.

Biochemical Reactions The behaviour of a MoK system is actually determined by the last abstraction of the model: the *biochemical reaction* [1].

As a knowledge-oriented model, the main issue of MoK is determining the semantic correlation between MoK atoms. So, to define a working MoK system, the basic $mok(\ atom_1,\ atom_2\ )$ function should be defined, which takes two atoms $atom_1$, $atom_2$, and returns a *"matching degree"*—which could be a boolean or a double value $\in [0, 1]$. The precise definition of *mok* depends on the specific application domain.

The *aggregation reaction* (AggR) bonds together atoms and molecules:

$$molecule(Atoms_1)\ +\ molecule(Atoms_2) \longmapsto^{r\text{-}agg} molecule(Atoms_1 \bigcup Atoms_2)\ +\ Residual(Atoms_1, Atoms_2)$$

where r_agg is the AggR reaction rate, $mok(atom_1, atom_2)$ holds for some $atom_1 \in Atoms_1$, $atom_2 \in Atoms_2$, and $Residual(Atoms_1, Atoms_2)$ is the multiset of atoms obtained as the multiset difference $(Atoms_1 \uplus Atoms_2) \backslash (Atoms_1 \bigcup Atoms_2)$—that is, essentially, all the atoms of $Atoms_1$ and $Atoms_2$ that do not belong to the resulting molecule. In short, more complex molecules are formed by aggregation whenever some atoms in the reacting molecules (or in reacting atoms, handled here as one atom molecules) are semantically correlated (via the *mok* function).

Positive feedback is obtained by the *reinforcement reaction* (ReinfR), which consumes a single unit of enzyme to produce a unit of the relative atom/molecule:

$$enzyme(Atoms_1) \; + \; molecule(Atoms_2)_c \longmapsto^{r\_reinf} molecule(Atoms_2)_{c+1}$$

where r_reinf is the ReinfR reaction rate, $enzyme(Atoms_1)$ and $molecule(Atoms_2)_c$ exist in the compartment, both with $c \neq 0$, and $mok(atom_1, atom_2)$ holds for some $atom_1 \in Atoms_1$, $atom_2 \in Atoms_2$

Following the biochemical metaphor, molecules should fade as time passes, lowering their own concentration according to some well-defined *decay-law*. The temporal *decay reaction* (DecayR) is hence defined as follows:

$$molecule(Atoms)_c \longmapsto^{r\_decay} molecule(Atoms)_{c-1}$$

Similarly, a biochemical-inspired knowledge management model should provide some spatial interaction pattern. MoK adopts *diffusion* as its data-migration mechanism, in which atoms and molecules can move only between *neighbour* compartments, resembling membrane crossing among cells. MoK *diffusion reaction* (DiffR) is then modelled as follows, assuming that σ identifies a biochemical compartment, and $\|\|_\sigma$ brackets molecules in a compartment σ:

$$\begin{array}{c} \| \; Molecules_1 \bigcup molecule_1 \; \|_{\sigma^i} \; + \| \; Molecules_2 \; \|_{\sigma^{ii}} \longmapsto^{r\_diffusion} \\ \| \; Molecules_1 \; \|_{\sigma^i} \; + \| \; Molecules_2 \bigcup molecule_1 \; \|_{\sigma^{ii}} \end{array}$$

where σ^i and σ^{ii} are neighbour compartments, $r\_diffusion$ is the diffusion rate, and $molecule_1$ moves from σ^i to σ^{ii} as the result of the reaction.

2 News Management: A First Case Study

While MoK is a general-purpose model for knowledge self-organisation, it can be tailored on specialised application scenarios, by refining the notions of atom and suitably defining the *mok* semantic correlation function. Since news management provide a prominent example of a knowledge-intensive environment, we chose news management as the first case study for MoK, introducing the MoK-News model for self-organisation of news.

2.1 Knowledge representation for news management

IPTC[1] develops and maintains technical standards for improved news management, such as NewsML[2] and NITF[3].

[1] `http://www.iptc.org/`
[2] `http://www.iptc.org/site/News_Exchange_Formats/NewsML-G2/`
[3] `http://www.iptc.org/site/News_Exchange_Formats/NITF/`

The NewsML tagging language is a media-type orthogonal news sharing format standard aimed at conveying not only the core news *content*, but also the data that describe the content in an abstract way, namely the *metadata*. In order to ease syntactical and semantical interoperability, NewsML adopts XML as the first implementation language for its standards and for maintaining sets of *Controlled Vocabularies* (CVs), collectively branded as *NewsCodes*, to represent concepts describing and categorising news objects in a consistent manner—pretty much as domain-specific ontologies do.

The News Industry Text Format, too, adopts XML to enrich the content of news articles, supporting the identification and description of a number of news typical features, among which the most notable are: *Who* owns the copyright to the item, who may republish it, and who it's about; *What* subjects, organisations, and events it covers; *When* it happened, was reported, issued, and revised; *Where* it was written, where the action took place, and where it may be released; *Why* it is newsworthy, based on the editor's analysis of the metadata. NewsML in fact provides no support for any form of inline tagging to add information to the plain text, for instance with the purpose to ease the work of a text mining algorithm usable to automatically process the document. Thus, NITF and NewsML are complementary standards, hence they perfectly combine to shape quite a comprehensive and coherent framework to manage the whole news lifecycle: comprehensive, given that one cares about news overall structure, including metadata, whereas the other focusses on their internal meaning making it unambiguous; coherent, because they both exploit the same IPTC abstractions—in fact NITF, too, uses NewsCodes.

2.2 Towards MoK-News

Since sources provide journalist with the required *raw information* already formatted according to the afore-mentioned IPTC standards, a simple-yet-effective mapping can be drawn. In fact, MoK atoms do actually have a clear counterpart in NewsML and NITF standards: *tag*. Tags – along with their "content" – can in fact be seen as the atoms that altogether compose the "news-substance" in the news management scenario. As a result, our MoK-coordinated news management system would contain $< newsItem >$ atoms, $< person >$ atoms, $< subject >$ atoms, etc.—that is, virtually one kind of atom for each NewsML/NITF tag.

A MoK-News atom looks like $atom(\ src,\ val,\ sem(tag, catalog)\ )_c$, where

$$
\begin{aligned}
&src\ ::=\ news\ source\ uri\\
&val\ ::=\ news\ content\\
&attr\ ::=\ sem(tag,\ catalog)\\
&\qquad tag\ ::=\ NewsML\ tag\ |\ NITF\ tag\\
&\qquad catalog\ ::=\ NewsCode\ uri\ |\ ontology\ uri
\end{aligned}
$$

Here, the content of an atom is mostly given by the pair $< val, tag >$, where *tag* could be either a metadata tag drawn from NewsML or an inline description tag taken from NITF. The precise and unambiguous semantics of the new content (*val*) can be specified thanks to the *catalog* information, which could be grounded in either NewsML or NITF standards in the form of NewsCodes, or

instead be referred to a custom ontology defined by the news worker. MoK-News molecules, enzymes, and biochemical reactions are then both syntactically and semantically affected by such domain-specific mapping of the MoK model.

3 Related & Future Works

To the best of our knowledge, althought the (bio-)chemical metaphor is widely used to achieve self-organisation and self-adaptation by emergence, no MoK-like approaches were studied that could bring self-* behaviours directly *into data*. On the other hand, in the news community much attention is paid to interoperability and semantic standardisation—but no *paradigm shift* has been tried to "see" data as active entities.

Nevertheless, some MoK-similar system actually exists although with different aims. In [2] a general-purpose, tuple-space-based approach to knowledge self-organisation was built exploiting a WordNet ontology to identify relationships between knowledge chunks – tuples –, and drive their migration to build clusters. In [3] a similar clustering behavior is achieved with *collective sort*, assuming that a $1:1$ relation exists between admissible tuple templates and tuple "sorts"—essentially making the number of clusters known *a priori*. The MoK approach is different in that it pushes the above cited "similarity-based clustering" to the limit: MoK not only aggregates somehow-related knowledge chunks in a same spot – e.g. diffusing news to interested journalists – but also tries to *physically merge* units of information to create new knowledge—the molecules.

Our future efforts will be on first devoted to provide an effective implementation of the MoK model upon an existing coordination middleware enriched with the "online biochemical simulator" behavior [4], then to test the implementation on the MoK-News application scenarios, and on other knowledge-intensive environments as well—e.g., *MoK-Research* and *MoK-HealthCare*.

References

1. Viroli, M., Casadei, M., Nardini, E., Omicini, A.: Towards a chemical-inspired infrastructure for self-* pervasive applications. In Weyns, D., Malek, S., de Lemos, R., Andersson, J., eds.: Self-Organizing Architectures. Volume 6090 of LNCS. Springer (July 2010) 152–176
2. Pianini, D., Virruso, S., Menezes, R., Omicini, A., Viroli, M.: Self organization in coordination systems using a WordNet-based ontology. In Gupta, I., Hassas, S., Jerome, R., eds.: 4th IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO 2010), Budapest, Hungary, IEEE CS (27 September–1 October 2010) 114–123
3. Gardelli, L., Viroli, M., Casadei, M., Omicini, A.: Designing self-organising MAS environments: The collective sort case. In Weyns, D., Parunak, H.V.D., Michel, F., eds.: Environments for MultiAgent Systems III. Volume 4389 of LNAI. Springer (May 2007) 254–271
4. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry **81**(25) (1977) 2340–2361

Semantic Models for Question Answering

Piero Molino
Supervisor: Pasquale Lops

Dept. of Computer Science, University of Bari
Via Orabona – I-70125 Bari, Italy
piero.molino@uniba.it

Abstract. The research presented in this paper focuses on the adoption of semantic models for *Question Answering* (QA) systems. We propose a framework which exploits semantic technologies to analyze the question, retrieve and rank relevant passages. It exploits: (a) *Natural Language Processing* algorithms for the analysis of questions and candidate answers both in English and Italian; (b) *Information Retrieval* (IR) probabilistic models for retrieving candidate answers and (c) *Machine Learning* methods for question classification. The data source for the answers is an unstructured text document collection stored in search indices. The aim of the research is to improve the system performances by introducing semantic models in every step of the answering process.

1 Introduction

Question Answering (QA) is the task of answering users' questions with answers obtained from a collection of documents or from the Web.

Traditional search engines usually retrieve long lists of full-text documents that must be checked by the user in order to find the needed information. Instead QA systems exploit *Information Retrieval* (IR) and *Natural Language Processing* (NLP) [18, 9], to find the answer, or short passages of text containing it, to a natural language question. Open-domain QA systems search on the Web and exploit redundancy, textual pattern extraction and matching to solve the problem [14, 12].

QA emerged in the last decade as one of the most promising fields in *Artificial Intelligence* thanks to some competitions organized during international conferences [19, 16], but the first studies on the subject can be dated back to 1960s [3]. In the last years some enterprise applications, such as IBM's Watson/DeepQA [7], have shown the potential of the state-of-the-art techology.

This paper describes a study on the introduction of semantic models inside a QA framework in order to improve its performances in answering users' questions. Although Semantic Role Labelling and Word Sense Disambiguation have been already employed in the past [18], distributional and latent models for QA are a completely new approach to investigate for QA.

A framework for building real-time QA systems with focus on closed domains was built for this purpose. The generality of the framework allows that also its

application to open domains can be rather easy. It exploits NLP algorithms for both English and Italian and integrates a question categorization component based on Machine Learning techniques and linguistic rules written by human experts. Text document collections used as data sources are organized in indices for generic unstructured data storage with fast and reliable search functions exploiting state-of-the-art IR weighting schemes.

The paper is structured as follows. Section 2 provides a generic overview of the framework architecture, while Section 3 presents the different semantic models. In Section 4 a preliminary evaluation of the impact of the adoption of semantic models is provided. Final conclusions, then, close the paper.

2 Framework overview

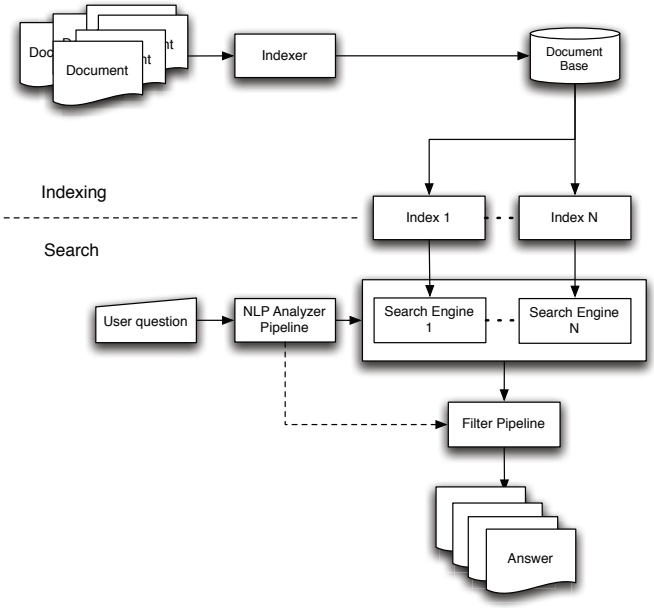

Fig. 1. QC1 Framework architecture overview

The architecture, shown in Figure 1, introduces some new aspects that make it general and easier to expand, such as the adoption of different indices, parallel search engines and different NLP and filtering pipelines, which can also run in parallel.

The first step is a linguistic analysis of the user's question. Question analysis is performed by a pipeline of NLP analyzers. NLP analyzers are provided both for

English and Italian and include a stemmer, a Part of Speech tagger, a lemmatizer, a Named Entity Recognizer and a chunker.

This step includes also a question classifier that uses an ensemble learning approach exploiting both hand-written rules and rules inferred by machine learning categorization techniques (*Support Vector Machines* are adopted), thus bringing together the hand-written rules' effectiveness and precision and the machine learning classifier's recall.

The question is then passed to the search engines, whose architecture is highly parallel and distributed. Moreover, each single engine has its own query generator, because the query's structure and syntax could change between different engines. For this purpose two different query expansion techniques are implemented: *Kullback-Liebler Divergence* [4] and *Divergence From Randomness* [1].

The filter pipeline is then responsible for the scoring and filtering of the passages retrieved by the search engines. Finally, a ranked list of passages is presented to the user.

3 Semantic Models

The word "semantics" describes the study of meaning. In NLP and IR it is used to refer to "lexical semantics", i.e. the meaning of words, and to "semantic role", the role of a phrase in a sentence. The aim of this research is to investigate whether semantic models can improve performances of QA systems and under which conditions improvements are achieved. A deep cost-benefit analysis analysis of semantic models will be also performed. New models will also be developed and tested.

From the point of view of QA, semantic models can be applied to different parts of the process.

Among the NLP Analyzers, *Word Sense Disambiguation* techniques should be applied to find the explicit meaning of every word taken from a semantic lexicon like *WordNet* [6]. The application of *Semantic Role Labelling* algorithms is also needed to extract the role of a phrase, thus helping the identification of the most important parts of the user's question [18]. *Word Sense Induction* [13] methods should be also applied to discriminate word meaning depending on the use of the word inside the collection, in an implicit way.

In the query expansion step, the use of synonyms taken from explicit repositories and similar words obtained from the similarity of contexts of use need both to be investigated.

During the search step, different approaches to semantic analysis should be applied, ranging from algebraic matrix approaches like *Latent Semantic Analysis* [5], *Non-negative Matrix Factorization* [11] and *Random Indexing* [10], to explicit representation approaches like *Explicit Semantic Analysis* [8]. As most of these techniques are expensive to be applied in real time for all documents, their adoption can be shifted to the filtering step, thus applying them only on a reduced and pre-filtered subset of all the candidate answers. To the best of my knowledge those semantic models have never been adopted in QA, in particular

for candidate answer filtering and scoring. Different search models like fuzzy and neural network based models for IR [2] will also be investigated.

In the filtering step, semantic distance and semantic correlation [15] measures can be applied to score the candidate answers, according to the adopted representation of meaning.

4 Preliminary Evaluation

A preliminary evaluation has been conducted on the *ResPubliQA 2010 Dataset* adopted int he *2010 CLEF QA Competition* [16]. This dataset contains about 10700 documents about European Union Legislation and European Parliament transcriptions, aligned in several languages including English and Italian, with 200 questions.

The adopted metric is the *c@1* proposed for the competition:

$$c@1 = \frac{1}{N}\left(n_c + n_n \frac{n_c}{N}\right) \tag{1}$$

where N is the number of the questions, n_c is the number of the system's correct answers and n_n is the number of unanswered questions.

Several combinations of different parameters have been tested, but in the table below only the best one is shown. It employs a two BM25 based searchers [17] as search component, one using a keyword index, the other using a lemma index, and the filters pipeline is built with keyword matching, lemma matching, exact overlap, density and n-grams filters. The framework is named **qc1**.The adoption of a *Random Indexing* based semantic filter improves the system performance of 0.045 for English and 0.04 for Italian, as shown in Table 1.

System	Search	k1	b	RI	c@1 score
qc1 english	BM25	1.6	0.8	no	0.705
qc1+sf english	BM25	1.6	0.8	yes	0.75
best CLEF2010 en					0.73
qc1 italian	BM25	1.8	0.75	no	0.635
qc1+sf italian	BM25	1.8	0.75	yes	0.675
best CLEF2010 it					0.63

Table 1. Preliminary evaluation

5 Conclusions

In this paper, a research proposal about the adoption of semantic models for QA has been presented. A short overview of the adopted QA framework, alongside with a description of the different semantic models to adopt for the research purpose, has been provided Finally, a preliminary evaluation on a standard dataset,

CLEF 2010 ResPubliQA, has been given, which shows an improvement in comparison to other state-of-the-art systems and demonstrates how promising the use of semantic models is for the field of QA.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20, 357–389 (October 2002)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. Paperback (May 1999)
3. Bert F. Green, J., Wolf, A.K., Chomsky, C., Laughery, K.: Baseball: an automatic question-answerer. In: Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. pp. 219–224. IRE-AIEE-ACM '61 (Western), ACM, New York, NY, USA (1961)
4. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Trans. Inf. Syst. 19, 1–27 (January 2001)
5. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
6. Fellbaum, C.: WordNet: an electronic lexical database. Language, speech, and communication, MIT Press (1998)
7. Ferrucci, D.A.: Ibm's watson/deepqa. SIGARCH Computer Architecture News 39(3) (2011)
8. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: In Proceedings of the 20th International Joint Conference on Artificial Intelligence. pp. 1606–1611 (2007)
9. Hovy, E.H., Gerber, L., Hermjakob, U., Junk, M., Lin, C.Y.: Question answering in webclopedia. In: TREC (2000)
10. Kanerva, P.: Sparse distributed memory. Bradford Books, MIT Press (1988)
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (Oct 1999)
12. Lin, J.: An exploration of the principles underlying redundancy-based factoid question answering. ACM Trans. Inf. Syst. 25 (April 2007)
13. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2), 10:1–10:69 (Feb 2009)
14. Paşca, M.: Open-domain question answering from large text collections. Studies in computational linguistics, CSLI Publications (2003)
15. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity: measuring the relatedness of concepts. In: Demonstration Papers at HLT-NAACL 2004. pp. 38–41. HLT-NAACL–Demonstrations '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
16. Penas, A., Forner, P., Rodrigo, A., Sutcliffe, R.F.E., Forascu, C., Mota, C.: Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. In: Braschler, M., Harman, D., Pianta, E. (eds.) Working notes of ResPubliQA 2010 Lab at CLEF 2010 (2010)
17. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. 3, 333–389 (April 2009)

An ASP Approach for the Optimal Placement of the Isolation Valves in a Water Distribution System

Andrea Peano
Supervisor: Marco Gavanelli

EnDiF, Università degli Studi di Ferrara
via G. Saragat 1 – 44122 Ferrara, Italy
andrea.peano@unife.it

1 Introduction

My Ph.D. Thesis relates to real-life optimization problems in the hydraulic engineering field. More precisely, with the collaboration of computer scientists, operational researchers and hydraulic engineers, I investigate and exploit potentialities of various Operational Research and Artificial Intelligence techniques in order to achieve good (and, whenever possible, optimal) solutions for those particular design issues of the urban hydraulic network that can be effectively modelled as known combinatorial optimization problems. Furthermore, such design issues often require to devise new specialized variants of the known combinatorial optimization problems.

For example, the problem of minimizing the impact of a contamination in a hydraulic network can be seen, under opportune assumptions, as a variant of the well known *Multiple Traveling Salesman Problem* (MTSP); since the quality of feasible solutions must be computed through a burdensome hydraulic simulation, such optimization problem was addressed by us by means of several genetic algorithms [6]. In particular, in [6], we proposed a novel genetic encoding for the MTSP for which we defined new genetic crossover operators based on ad hoc mixed integer linear programming (sub-)optimizations, obtaining a hybrid genetic algorithm.

Another real-life combinatorial optimization problem which is a typical issue during the design of a hydraulic network is finding the optimal positioning of a limited number of isolation valves on the network. Up to now, we exploited two different technologies, following two independent approaches: in the first we modelled the above mentioned problem by means of a Bilevel (Mixed Integer) Linear Programming [3] formalization, discussed in [13]; I presented the study at the 3rd Student Conference on Operational Research (SCOR 2012). In the second approach, we addressed such optimization problem by defining several Answer Set Programming (ASP)[1, 11, 8] programs, discussed in [5]. Both the optimization approaches compute the globally optimum placement of the valves.

In the next section I will briefly describe the problem of the isolation valve placement on hydraulic networks and the ASP approach designed to solve it.

2 The Isolation Valves Location Problem

Water Distributions Systems (WDSs) are strategic urban infrastructures. Their planning is, in turn, a strategic task in terms of costs control and to assure a fair degree of reliability. For example, during the design of a water distribution network, one of the choices is the design of the isolation system. It is a real-life problem for hydraulic engineers, and in recent years it has been studied through computational methods in the *hydroinformatics* literature [9, 4].

A water distribution system has the main objective of providing water to homes and facilities that require it. The water distribution network can be thought of as a labelled indirected graph, in which the edges represent the pipes in the network. There is at least one special node that represents the source of water (node 1 in Figure 1), and the users' homes are connected to the edges. For each edge, we assume to have knowledge about the average amount of water (in litres per second) that is drawn by the users insisting on that edge (during the day); such value is the label associated to the edge, and it is called the users' *demand*.

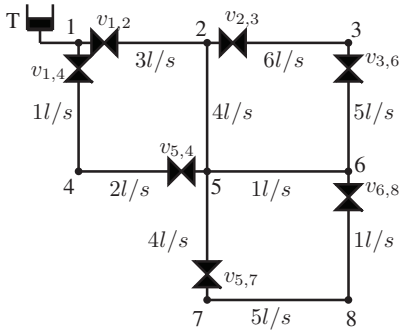

Fig. 1. A water distribution network with valves

The isolation system is mainly used during repair operations: in case some pipe is damaged, it has to be fixed or substituted. However, no repair work can be done while the water is flowing at high pressure in the pipe: first the part of the network containing the broken pipe should be de-watered, then workers can fix the pipe. The de-watering is performed by closing an opportune set of *isolation valves* (that make up the so-called *isolation system* of the water distribution network) so that the damaged pipe is disconnected from the sources. For example, in Figure 1, if the edge connecting nodes 2 and 3 (let us call it $e_{2,3}$) is broken, workers can close valves $v_{2,3}$ and $v_{3,6}$ and de-water the broken pipe. Of course, during this pipe substitution the users that take water from edge $e_{2,3}$ cannot be serviced. The usual measure of disruption is the *undelivered demand*: in this case, it corresponds to the demand of the users insisting on the broken pipe, i.e., $6l/s$.

However, we are not always this lucky: in case the damaged pipe is $e_{7,8}$, workers will have to close valves $v_{5,7}$ and $v_{6,8}$, de-watering pipes $e_{7,8}$ and $e_{6,8}$,

with a total cost of $5 + 1 = 6l/s$. In fact, the minimum set of pipes that will be de-watered is that belonging to the so-called *sector* of the broken pipe, i.e., the set of pipes encircled by a same set of valves. But there can be even worse situations: if the broken pipe is $e_{2,5}$, workers have to close valves $v_{1,2}$ and $v_{5,4}$, which means disconnecting all the pipes except $e_{1,4}$ and $e_{4,5}$, with an undelivered demand of $3+4+6+5+1+4+5+1 = 29l/s$. Notice in particular that the edges $e_{2,3}$, $e_{7,8}$ and $e_{6,8}$ are disconnected in this way, although they do not belong to the same sector as the broken pipe. This effect is called *unintended isolation*, and usually means that the isolation system was poorly designed.

One common value used by hydraulic engineers [9] to measure the quality of the isolation system is the undelivered demand in the worst case. In the example of Figure 1, the worst case happens when the broken pipe is in the set $\{e_{1,2}, e_{2,5}, e_{5,6}, e_{5,7}\}$, and the related suppy disruption is $29l/s$, as above.

In a previous work, [2] developed a system, based on Constraint Logic Programming [10] on Finite Domains (CLP(FD)), that finds the optimal positioning of a given number of valves in a water distribution network. The assignments found by [2] improved the state-of-the-art in hydraulic engineering for this problem, finding solutions with a lower (worst-case) undelivered demand than the best solutions known in the literature of hydraulic engineering [9], obtained through genetic algorithms.

In the current work, we address the same problem in Answer Set Programming [1, 11, 8], which is a suitable technology to address combinatorial graph problems[12], and, in particular, we have already defined two different ASP programs [5]. One program explicitly defines the *sectors* as clusters of (isolated) pipes and minimizes the undelivered demand of the worst sector; instead, in the other program, sectors are left implicit and the aim is to maximize the minimum satisfied demand in case of pipe isolation, by considering that a pipe is isolated if it is not reachable from any source. In the next section we show the most important results obtained by first experiments.

3 Results

The first experiments, presented in [5], show, in general, that the developed programs take more computation time than the CLP(FD) approach [2]. However, we must say that the CLP(FD) model was developed by two CLP experts, during some person-months and was trimmed for efficiency. Instead, the two ASP formulations were mainly developed by a first-year PhD student in about one week; this shows that ASP is very intuitive and easy to understand even for non experts, that it is indeed very declarative. The two implemented ASP programs consist of respectively about 20 and 25 rules, which shows that ASP is a very interesting technology for rapid prototyping.

Experiments have been performed on a *Intel* based architecture with two P8400 CPUs; as ASP solver we used the Potassco's solver *Clasp* [7]. The two programs have been optimized using a real-life instance based on the Apulian hydraulic network [9] (Figure 2) and varying the number of available isolation valves.

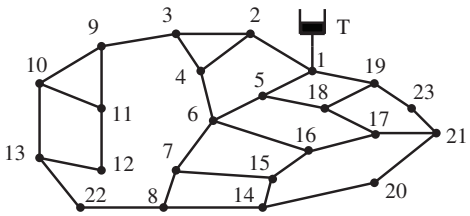

Fig. 2. The Apulian hydraulic network

Figure 3 shows the optimization performance of the two ASP programs (the one based on sectors and the one that, instead, does not define sectors) for several number of available valves. In particular, the better effectiveness of the sector-based program can be noticed.

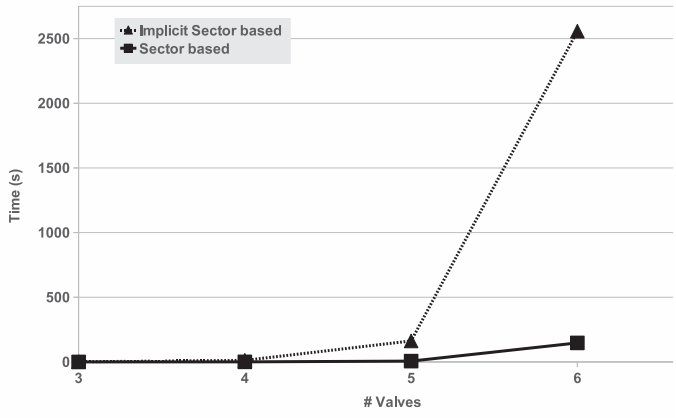

Fig. 3. Computing times of the optimization processes of the two ASP programs

4 Future Work

In future work, we plan to improve the sector-based ASP program by defining opportune rules in order to break the symmetries determined by the current formalization, and we will experiment the resulting program also with other available ASP solvers. Another appealing challenge is to compute the solution that minimizes the undelivered demand of the worst sector as well as the undelivered demands of the other (no worst) sectors, which are not actually optimized neither by our current MILP formalization [13] nor by the CLP(FD) one [2]. We are also interested in trying to integrate the ASP programs with a CLP approach, to take advantage of the strengths of the two approaches. Finally, I plan to delve into the ASP theory and techniques in order to consolidate my competence in such Artificial Intelligence field.

Acknowledgements This work was partially supported by EU project *ePolicy*, FP7-ICT-2011-7, grant agreement 288147. Possible inaccuracies of information are under the responsibility of the project team. The text reflects solely the views of its authors. The European Commission is not liable for any use that may be made of the information contained in this paper.

References

1. Baral, C.: Knowledge representation, reasoning and declarative problem solving. Cambridge University Press (2003)
2. Cattafi, M., Gavanelli, M., Nonato, M., Alvisi, S., Franchini, M.: Optimal placement of valves in a water distribution network with CLP(FD). Theory and Practice of Logic Programming 11(4-5), 731–747 (2011), http://arxiv.org/abs/1109.1248
3. Colson, B., Marcotte, P., Savard, G.: Bilevel programming: A survey. 4OR: A Quarterly Journal of Operations Research 3, 87–107 (2005)
4. Creaco, E., Franchini, M., Alvisi, S.: Optimal placement of isolation valves in water distribution systems based on valve cost and weighted average demand shortfall. Water Resources Management 24, 4317–4338 (2010)
5. Gavanelli, M., Nonato, M., Peano, A., Alvisi, S., Franchini, M.: An ASP approach for the valves positioning optimization in a water distribution system. In: Lisi, F. (ed.) 9th Italian Convention on Computational Logic (CILC 2012), Rome, Italy. CEUR workshop proceedings, vol. 857, pp. 134–148 (2012)
6. Gavanelli, M., Nonato, M., Peano, A., Alvisi, S., Franchini, M.: Genetic algorithms for scheduling devices operation in a water distribution system in response to contamination events. In: Hao, J.K., Middendorf, M. (eds.) Evolutionary Computation in Combinatorial Optimization, Lecture Notes in Computer Science, vol. 7245, pp. 124–135. Springer Berlin / Heidelberg (2012)
7. Gebser, M., Kaminski, R., Kaufmann, B., Ostrowski, M., Schaub, T., Schneider, M.: Potassco: The Potsdam answer set solving collection. AI Communications 24(2), 105–124 (2011)
8. Gelfond, M.: Answer sets. In: Handbook of Knowledge Representation, chap. 7. Elsevier (2007)
9. Giustolisi, O., Savić, D.A.: Identification of segments and optimal isolation valve system design in water distribution networks. Urban Water Journal 7(1), 1–15 (2010)
10. Jaffar, J., Maher, M.J.: Constraint logic programming: A survey. J. Log. Program. 19/20, 503–581 (1994)
11. Leone, N.: Logic programming and nonmonotonic reasoning: From theory to systems and applications. In: Baral, C., Brewka, G., Schlipf, J. (eds.) Proceedings of the 9th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'07), Lecture Notes in Computer Science, vol. 4483. Springer (2007)
12. Niemelä, I.: Logic programs with stable model semantics as a constraint programming paradigm. Annals of Mathematics and Artificial Intelligence 25, 241–273 (1999)
13. Peano, A., Nonato, M., Gavanelli, M., Alvisi, S., Franchini, M.: A Bilevel Mixed Integer Linear Programming Model for Valves Location in Water Distribution Systems. In: Ravizza, S., Holborn, P. (eds.) 3rd Student Conference on Operational Research. OpenAccess Series in Informatics (OASIcs), vol. 22, pp. 103–112. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2012)

Machine Learning Methods and Technologies for Ubiquitous Computing

Agnese Pinto
Supervisors: Eugenio Di Sciascio, Michele Ruta

DEE, Politecnico di Bari
via Re David 200 – I-70125 Bari, Italy
agnese.pinto@poliba.it

Abstract. The research proposal concerns the use of machine learning techniques for data mining in pervasive environments. It will lead to the formalization of a framework, able to translate series of "raw" data in high-level knowledge. Novel machine learning approaches, interpreting data coming from the environment that surrounds users, will be liveraged. Data will be collected through micro-components deployed in the field and will be processed for the identification and characterization of phenomena and contexts. Eventually they will be semantically annotated to support futher application-level logic-based reasoning and knowledge discovery.

1 Scenario

The Ubiquitous Computing paradigm, introduced by Mark Weiser [16], refers to manifold aspects involving pervasiveness in information storage, processing and discovery. It is aimed to a model of human-computer interaction where information as well as computational capabilities are deeply "integrated" into everyday objects and/or actions. In such vision, the increasing "availability" of processing power would be accompanied by its decreasing "visibility". As opposed to classical paradigms, where a user explicitly engages a single device to perform a specific task, exploiting ubiquitous computing features the user will interact with many computational devices simultaneously. She will extract data from the objects permeating the environment during ordinary activities, even not necessarily being aware of what is happening. By embedding short-range mobile transceivers into a wide array of devices and everyday objects, new kinds of interactions can be enabled between people and things, as well as between things themselves. This is the Internet of Things (IoT) vision [12]. The IoT paradigm is based on the use of a large number of heterogeneous micro devices, each conveying a small amount of useful information. Several emerging technologies are suitable to bridge the gap between physical things and the digital world. For instance, Radio Frequency IDentification (RFID) [7] allows object identification by means of electronic transponders (tags) attached to items. Wireless Sensor Networks (WSNs), on the other hand, allow to monitor environmental parameters, supporting queries and automatic alerts triggered by application-defined

events [9]. Both these technologies are characterized by the dissemination of unobtrusive, inexpensive and disposable micro-devices in a given environment.

From the user's standpoint, the goal of pervasive computing is to reduce the amount of user effort and attention required to benefit from computing systems. Current mobile resource discovery protocols have been directly derived from the ones originally designed for infrastructure-based wired networks. In fact, current technologies such as RFID, Bluetooth [2] and ZigBee [18] only allow string matching for item identification. Nevertheless, purely syntactic match mechanisms cannot support more advanced wireless applications, since they provide only boolean yes/no outcomes. It is desirable to manage requests and service descriptions with richer and unambiguous meaning, by adopting formalisms with well-grounded semantics. Wireless communication technologies and mobile computing systems are approaching sufficient maturity to overcome the above limitations. In particular, an advanced mobile resource discovery facility should be able to support non-exact matches [3] and to provide a ranked list of discovered resources or services. This allows to satisfy a user request "to the best possible extent" whenever fully matching resources/services are not available. To do so, techniques for Knowledge Representation (KR) and semantic-based matchmaking may be useful [6]. Hence the goal of pervasive knowledge-based systems is to embed semantically rich and easily accessible information into the physical world.

Nevertheless, in order to cope with pervasive computing contraints, techniques for objects and phenomena characterization are strongly needed. In fact, the context characterization is a foundamental issue in the semantic annotation of a scenario which is mandatory for further dicovery stages.

2 Proposal

Design and impelmentation of effective pervasive computing frameworks involve the study and testing of new technologies. Such technologies should allow to interpret in an automatic or semi-automatic way data extracted from the objects dipped in a generic context, traslating them in knowledge useful to automate processes or support user decisions and activities. In pervasive computing context, data often exhibit a high level complexity (different sensing/collecting technologies, huge volumes, inter-dependency relationships between sources) and dinamicity (real-time update and critical aging). There is the need for methods and algorithms characterized by accuracy, precision and timeliness, also taking into account the limits - in terms of computing resources, memory and communication - of devices.

This research proposal therefore aims to study and formalize methodologies and tools for managing data streams; data collected from a large number of micro-devices in specific environments, using data mining techniques, in particular machine learning algorithms. The semantic-based technologies for element annotation combined with Machine Learning (ML) theory can lead to the formalization of innovative frameworks for pervasive computing, able to trans-

form field data streams in knowledge. Such knowledge will be then usable in application-level decisional processes, to improve the human activities. In other words, raw data will be gathered through a series of micro-components deployed and dipped in a given environments; the collected data will be processed for the identification and characterization of phenomena and contexts. The phenomena and contexts so identified will be described through semantic annotations, in order to be processed by automated logic-based reasoning algorithms.

Data Mining is the process of discovering interesting hidden knowledge by identifying patterns from large amounts of data, where the data can be stored in databases, data warehouses or other information repositories. ML is an Artificial Intelligence area that provides the technical basis of Data Mining. ML manages an abstraction process which happens taking the data and inferring whatever structure underlies them [17], *i.e.*, the main goal of research in machine learning is to "learn" to automatically recognize complex patterns and to make intelligent decisions based on data already analyzed. A peculiarity of ML is *induction*, the extraction of general laws from a set of observed data. It is opposed to *deduction* where, starting from the general laws, the expected value of a set of variables is determined. Induction begins with the observation aiming at measuring a set of variables and then make predictions for further information. This process is named inference.

To carry out the proposed research we will proceed to identify the mainly used methods, with reference to mobile and pervasive computing environments to highlight both strengths and limits. We then will proceed to the definition of innovative annotation methodologies, through the integration and adaptation of existing technologies and/or the definition of new models, algorithms and techniques for machine learning and data analysis. Finally we will apply research outcomes in selected case studies under different scenarios.

3 Applications

Several approaches already exist devoted to pattern recognition to identify and characterize phenomena, events or activities of users in pervasive environments [1][8][11][4]. However, they have significant limitations in the characterization phase and in the support to the discovery of information resources. They are unable to characterize - al least partially - situations that do not correspond exactly to the rules established in the learning and system configuration phase. There are few approaches in literature that support advanced reasoning for mobile devices, which use inference algorithms for discovery, ranking and explanation of retrieval results.

Pervasive computing has many potential applications, from intelligent workplaces and smart homes to healthcare, gaming, leisure systems and to public transportation [10][13][5]. The approach and studies previously sketched could be applied in most of them.

In Wireless Semantic Sensor and Actor Networks (WSSANs) scenarios, using semantic-based techniques at the application layer, it is possible to define

Wireless Semantic Sensor Networks (WSSN) able to offer more versatile services than traditional WSN [14]. For example in a scene of danger, thanks to a set of sensors for the detection of environmental conditions, it is possible to interpret detected data using machine learning techiques and identify one or more events. Recognized events can be semantically annotated and such semantic descriptions can be employed for making inferences.

In infomobility and driving assistance context, for example, several frameworks hare been proposed that aim to improve the capabilities of navigation systems (see as an instance [15]). It provides an approach that could be futher used to retrieve and deliver information to regulate vehicular traffic and to improve the safety of travel. In these scenarios, the contribution of data mining is essential to extract high-level information from a large number of low-level parameters that can be scanned from the car (through the on-board diagnostics OBD protocol - http://www.arb.ca.gov/msprog/obdprog/obdprog.htm) and built-in micro-components of a smartphone device, to accurately characterize the system (vehicle + driver + environment) and to improve driving safety and efficiency.

4 Conclusion

The presented research proposal refers to the extraction of knowledge from pervasive contexts using ML technologis and as Weiser said: "ubiquitous computers will help overcome the problem of information overload. There is more information available at our fingertips during a walk in the woods than in any computer system, yet people find a walk among trees relaxing and computers frustrating. Machines that fit the human environment instead of forcing humans to enter theirs will make using a computer as refreshing as taking a walk in the woods" [16].

References

1. Agarwal, I., Krishnaswamy, S., Gaber, M.M.: Resource-aware ubiquitous data stream querying. In: International Conference on Information and Automation (2005)
2. Bluetooth: http://www.bluetooth.com
3. Chakraborty, D., Perich, F., Avancha, S., Joshi, A.: Dreggie: Semantic service discovery for m-commerce applications. In: Workshop on Reliable and Secure Applications in Mobile Environment, 20th Symposiom on Reliable Distributed Systems (2001)
4. Choi, S., Kim, J., Kwak, D., Angkititrakul, P., Hansen, J.: Analysis and classification of driver behavior using in-vehicle can-bus information. In: Biennial Workshop on DSP for In-Vehicle and Mobile Systems. pp. 17–19. Citeseer (2007)
5. Chon, J., Cha, H.: Lifemap: A smartphone-based context provider for location-based services. Pervasive Computing, IEEE 10(2), 58–67 (2011)
6. Colucci, S., Di Noia, T., Pinto, A., Ruta, M., Ragone, A., Tinelli, E.: A nonmonotonic approach to semantic matchmaking and request refinement in e-marketplaces. International Journal of Electronic Commerce 12(2), 127–154 (2007)

7. Epcglobal Inc.: EPC Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Protocolfor Communications at 860 MHz–960 MHz. Tech. rep., EPCglobal (January 2005)
8. Fan, W., an Huang, Y., Wang, H., Yu, P.S.: Active mining of data streams. In: Fourth SIAM International Conference on Data Mining. pp. 457–461 (2004)
9. Gracanin, D., Eltoweissy, M., Wadaa, A., DaSilva, L.A.: A service-centric model for wireless sensor networks. IEEE Journal on Selected Areas in Communications 23(6), 1159–1166 (2005)
10. Hagras, H., Callaghan, V., Colley, M., Clarke, G., Pounds-Cornish, A., Duman, H.: Creating an ambient-intelligence environment using embedded agents. IEEE Intelligent Systems 19(6), 12–20 (Nov 2004), http://dx.doi.org/10.1109/MIS.2004.61
11. Hina, M., Tadj, C., Ramdane-Cherif, A.: Machine learning-assisted device selection in a context-sensitive ubiquitous multimodal multimedia computing system. In: Industrial Electronics, 2006 IEEE International Symposium on. vol. 4, pp. 3014–3019. IEEE (2006)
12. ITU: Internet Reports 2005: The Internet of Things. Tech. rep., ITU (November 2005)
13. Noguchi, K., Somwong, P., Matsubara, T., Nakauchi, Y.: Human intention detection and activity support system for ubiquitous autonomy. In: Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on. vol. 2, pp. 906–911. IEEE (2003)
14. Ruta, M., Scioscia, F., Di Noia, T., Di Sciascio, E.: A hybrid zigbee/bluetooth approach to mobile semantic grids. Computer Systems Science and Engineering 25(3), 235 (2010)
15. Ruta, M., Scioscia, F., Ieva, S., Di Sciascio, E.: T$\hat{2}$.o.m. t.o.m.: Techniques and technologies for an ontology-based mobility tool with open maps. Current Trends in Web Engineering pp. 199–210 (2010)
16. Weiser, M.: The computer for the 21st century. ACM SIGMOBILE Mobile Computing and Communications Review 3(3), 3–11 (1999)
17. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Amsterdam, 3. edn. (2011)
18. ZigBee Alliance: ZigBee Specification. ZigBee Alliance, 053474r17 edn. (January 17 2008)

Combining Process and Ontological Modeling

Dmitry Solomakhin
Supervisors: Sergio Tessaris, Marco Montali

Faculty of Computer Science, Free University of Bozen-Bolzano
Piazza Domenicani 3 – 39100 Bolzano, Italy
solomakhin@inf.unibz.it

1 Introduction and Motivation

Recent development of information technology has significantly affected the way how an enterprize operate. Nowadays in corporate information systems not only process automation but also dealing with all stages of the business process lifecycle becomes increasingly important. Relevant tasks cover not only issues to be tackled on the design phase (e.g. process modeling, simulation, verification of different properties of a process, etc) but also problems arising in the phases of execution and analysis (such as process mining, monitoring, etc). All these tasks are the central issues of Business Process Management (BPM).

In order to enable automated reasoning support for processes along their entire lifecycle, several mathematical formalisms have been adopted to represent processes as formal models, including transition systems, process algebras and, finally, Petri nets. However, the well-known and avowed disadvantage of the family of these approaches, often referred to as process-centric, is that although they capture the workflow of the process itself, most of them abstract away from the semantics of data which a process might operate with. Nowadays such data is usually of a very complex structure due to the nature of information to be described (e.g. logistics, sales, etc.). Therefore, dealing with such data requires powerful tools even for static analysis. Moreover, in real life business processes usually require data integration because data may originate from heterogeneous sources. With respect to the importance of data integration many process modeling methodologies do not provide appropriate conceptual paradigms for specifying and enacting these kinds of tasks [7].

In response to such a drawback of traditional process-centric BPM techniques, a data-centric business modeling has recently emerged as a methodology in which processes are considered to be driven with the possible changes and evolutions of business data objects, called *artifacts*. This approach has become an area of growing interest, since it has been argued that considering data-centric perspective in business modeling can lead to substantial cost savings in the design and deployment of business processes [2].

Following the current trend of knowledge-aware business process modeling, in our research we address a problem of merging the process-related modeling techniques with ontological modeling and semantic technologies in general. The final goal would be providing a logical/formal framework which allows for

modeling of business processes tightly coupled with the manipulated dynamic data, as well as for reasoning about and verification of different logical properties of such system. Such synthesis of models incorporating both a static and dynamic perspective, if exists, will require a very challenging task on defining algorithms for reasoning tasks, e.g. model checking, since in general that might lead to infinite-state models. Hence, not only new model checking algorithms have to be invented, but also decidable fragments of this combination should be investigated, mediating between relevance in practice and tractability [4].

2 The context of the research

The problems that are to be tackled along the research line can be considered relevant and useful in the context of the ACSI project [2], which is devoted to investigation on how the artifact-based approach may be used to optimize the business process management in the enterprize. The paradigm adopted there is presented in Figure 1 and consists of three layers: realization, artifact and semantics. The planned PhD research shares this paradigms and is supposed to focus on the semantic layer, i.e. to investigate the integration between the knowledge base describing the semantics of the data (ontologies) and high-level description of the business processes.

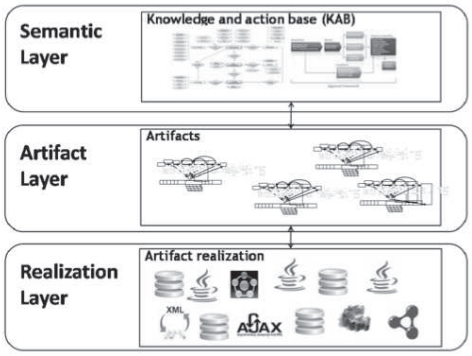

Fig. 1. ACSI Artifact Paradigm [2]

3 Current work

The recently introduced Guard-State-Milestone (GSM) artifact modeling language [6] provides means for specifying business artifacts lifecycles in a declarative manner, using intuitively natural constructs that correspond closely to

how business-level stakeholders think about their business. The corresponding constructs are:

- *Information model* for modeling relevant data domain.
- *Milestones*, which naturally correspond to business operational objectives and are achieved based on triggering events and/or conditions over the information model.
- *Stages*, which correspond to clusters of activities intended to achieve milestones and which can have a hierarchical structure.
- *Guards*, which control when a stage can be activated.

As an example, let's consider a process $Func$ which is as simple as calculating a square root of a sum $\sqrt{a+b}$, given that $a \neq b$ and $a+b \geq 0$. The GSM concrete model of such process is represented on the Figure 2.

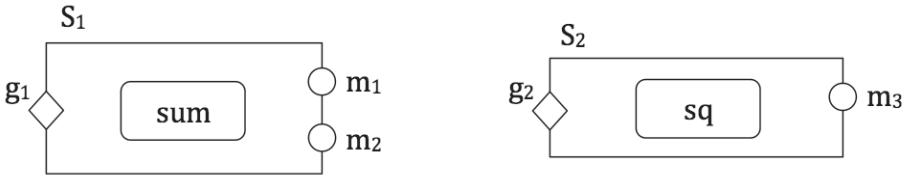

Fig. 2. GSM model of $\sqrt{a+b}$

Both milestones and guards are controlled in a declarative manner and corresponding definitions will have the following form:

$$\widetilde{g}_1 : \text{on } x.Func^{call}(a,b) \text{ if } a \neq b \qquad \widetilde{g}_2 : \text{on } +x.m_1$$
$$\widetilde{m}_1 : \text{on } x.Sum^{return}(c) \text{ if } c \geq 0 \qquad \widetilde{m}_2 : \text{if } c < 0$$
$$\widetilde{m}_3 : \text{on } x.Sq^{return}(d)$$

Despite having a formally specified operational semantics for GSM models [3], the verification of different properties of such models (e.g. existence of complete execution, safety properties) is still an open problem. In order to solve this problem, one should define a particular formalism that captures the intended operational semantics of the business artifacts and provides mechanisms to solve different verification tasks.

One of the most promising candidates for such a formalism is a data-centric dynamic system (DCDS) together with its general verification framework presented in [5]. A DCDS is a pair $\mathcal{S} = \langle \mathcal{D}, \mathcal{P} \rangle$, where $\mathcal{D}$ is a data layer and $\mathcal{P}$ is a process layer over the former.

The data layer $\mathcal{D}$ models the relevant database schema together with its set of integrity constraints, while the process layer $\mathcal{P}$ is a tuple $\mathcal{P} = \langle \mathcal{F}, \mathcal{A}, \varrho \rangle$, where

- $\mathcal{F}$ is a finite set of functions representing interfaces to external services.

- $\mathcal{A}$ is a set of actions of the specific form:

  $\alpha(p_1, ..., p_n) : \{e_1, ..., e_m\}$, where $p_1, ..., p_n$ are input parameters of an action and $e_i = q_i^+ \wedge Q_i^- \rightsquigarrow E_i$ are effects of an action of a particular form.

- ϱ is a process which is a finite set of condition-action rules of the form $Q \mapsto \alpha$, where α is an action and Q is a FO query over $\mathcal{R}$.

The decidability problem in the context of data-centric dynamic systems is one of the main challenges being investigated at the moment. However, unlike GSM, which has emerged to satisfy practical needs, DCDS benefits from having purely formal foundations, which provide instruments to approach and solve the challenge. Several decidability results have been obtained by Calvanese et. al [1] during their ongoing research. Therefore, it becomes of a particular interest to investigate the possibility to transfer these decidability results on GSM models.

Having a formal definition of an artifact and its lifecycle as a GSM concrete model, we aim to define a mapping which maps the artifact's relational schema into the data layer of DCDS and the set of ECA-like rules describing its behavior into the set of condition-action rules of the process layer of DCDS, where service calls are modeled by a finite set of functions $\mathcal{F}$.

Along the process of constructing the mapping we need to insure that the resulting DCDS model mimics the operational semantics of the initial GSM model. In particular, one would want to preserve the semantics of so-called B-Steps, which focus on what happens to a snapshot (i.e., description of all relevant aspects of a GSM system at a given moment of time) when a single incoming event is incorporated into it. In order to capture the semantics of B-Steps, we construct a so-called *conditional dependency graph*, which is then used to enforce the shape of the resulting condition-action rules in such a way that the final DCDS formalization may be, in fact, considered as an execution engine for the initial GSM model.

For example, assume a stage s_j and some guard $g_j^e = \textbf{on } \xi(x) \textbf{ if } \phi(x)$ which opens the stage. Then activating the stage by validating g_j^e can be modeled by the following condition-action rule:

$$\begin{aligned}
&\exists \overline{a}, \overline{s}, \overline{m}\ R_{att}(x, \overline{a}, \overline{s}, \overline{m}) \wedge s_j = false \wedge R_{Blocked}(x, false) \mapsto \\
&\qquad \alpha_{M,s_j}^{Activate}(id_R, a'_1, ..., a'_m) : \{ \\
&\qquad R_{att}(id_R, \overline{a}, \overline{s}, \overline{m}) \rightsquigarrow R_{att}(id_R, \overline{a}, \overline{s}, \overline{m})[s_j/true, a_1/f^M(1), ..., a_k/f^M(k)] \\
&\qquad R_{att}(id_R, \overline{a}, \overline{s}, \overline{m}) \rightsquigarrow R_M(id_R, f^M(1), ..., f^M(k)) \\
&\qquad R_{att}(id_R, \overline{a}, \overline{s}, \overline{m}) \rightsquigarrow R_{Block}(id_R, true) \\
&\qquad R_{att}(x, \overline{a}, \overline{s}, \overline{m}) \rightsquigarrow R_{s_j}^{StateChanged}(x, true)\}
\end{aligned}$$

4 Future work and concluding remarks

The results of the ongoing research are at the moment considered to be preliminary and subject to further investigation. In particular, one of the main future

tasks is verifying that the introduced translation from GSM model specification into DCDS specification is consistent with respect to a certain family of the process properties. This is going to be done by attempting to define a bisimulation relation between two transition systems, inferred by the semantics of GSM and DCDS respectively. Another task is devoted to investigating the possibility to transfer the existing decidability results for DCDS [1] to GSM and to determine expressivity restrictions corresponding to those defined for DCDS.

Other future tasks in the context of the ACSI project include: a) determining the use cases for the semantic layer in the ACSI Artifact paradigm, which tasks can be (or should be) dealt with on this layer; b) attempting to define a "semantic concrete model" which would be an "implementation" of a semantic layer of the ACSI Artifact Abstract Model, or more specifically, how to complement a GSM Concrete Model with some notion representing the semantic layer.

References

1. Bagheri-Hariri, B., Calvanese, D., De Giacomo, G., Deutsch, A., Montali, M.: Verification of relational data-centric dynamic systems with external services (2011), to appear
2. Calvanese, D., De Giacomo, G., Lembo, D.: The core ACSI artifact paradigm: artifact-layer and realization-layer. Public deliverable, The ACSI Project (FP7-ICT-2009-5-Objective 1.2, grant agreement 257593) (2011)
3. Damaggio, E., Hull, R., Vaculín, R.: On the equivalence of incremental and fixpoint semantics for business artifacts with guard-stage-milestone lifecycles. In: Proceedings of the 9th international conference on Business process management. pp. 396–412. BPM'11, Springer-Verlag, Berlin, Heidelberg (2011)
4. Hariri, B.B., Calvanese, D., De Giacomo, G., De Masellis, R., Felli, P.: Foundations of relational artifacts verification. In: Proceedings of the 9th international conference on Business process management. pp. 379–395. BPM'11 (2011)
5. Hariri, B.B., Calvanese, D., Giacomo, G.D., Masellis, R.D.: Verification of conjunctive-query based semantic artifacts. In: Rosati, R., Rudolph, S., Zakharyaschev, M. (eds.) Description Logics. CEUR Workshop Proceedings, vol. 745. CEUR-WS.org (2011)
6. Hull, R., Damaggio, E., De Masellis, R., Fournier, F., Gupta, M., Heath, III, F.T., Hobson, S., Linehan, M., Maradugu, S., Nigam, A., Sukaviriya, P.N., Vaculin, R.: Business artifacts with guard-stage-milestone lifecycles: managing artifact interactions with conditions and events. In: Proceedings of the 5th ACM international conference on Distributed event-based system. pp. 51–62. DEBS '11, ACM, New York, NY, USA (2011)
7. Volz, B.: Implementing conceptual data integration in process modeling methodologies for scientific applications. In: Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems: 2008 Workshops: ADI, AWeSoMe, COMBEK, EI2N, IWSSA, MONET, OnToContent + QSI, ORM, PerSys, RDDS, SEMELS, and SWWS. pp. 54–63 (2008)

Author Index