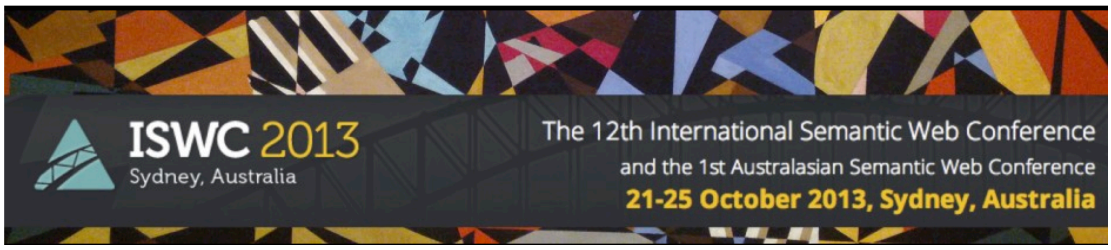
CrowdSem 2013

The Confluence of Crowdsourcing and Semantic Web



Doctoral Consortium

held at

the 12th International Semantic Web Conference

Sydney, Australia

Editors:

Lora Aroyo

Natasha Noy

Preface

This volume contains the papers presented at ISWC2013-DC: ISWC2013 Doctoral Consortium held on October 22, 2013 in Sydney.

There were 29 submissions. Each submission was reviewed by at least 2, and on the average 2.0, program committee members. The committee decided to accept 14 papers. Five of these papers were included in the main conference proceedings published by Springer. Nine papers are included in this volume.

We gratefully acknowledge the support University of Technology, Sydney, as the Doctoral Consortium sponsor.

October 22, 2013
Sydney, Australia

Natasha Noy
Lora Aroyo

Table of Contents

Program Committee

Lora Aroyo	VU University Amsterdam
Abraham Bernstein	University of Zurich
Oscar Corcho	Universidad Politécnica de Madrid
Mathieu D'Aquin	Knowledge Media Institute, the Open University
David Karger	MIT
Diana Maynard	University of Sheffield
Enrico Motta	Knowledge Media Institute, The Open University
Natasha F. Noy	Stanford University
Marta Sabou	MODUL University Vienna
Guus Schreiber	VU University Amsterdam
Elena Simperl	University of Southampton

Enriching Ontologies through Data

Mahsa Chitsaz*

School of Information and Communication Technology,
Griffith University, Australia
mahsa.chitsaz@griffithuni.edu.au

Abstract. Along with the vast usage of ontologies in different areas, non-standard reasoning tasks have started to emerge such as concept learning which aims to drive new concept definitions from given instance data of an ontology. This paper proposes new scalable approaches in light-weight description logics which rely on an inductive logic technique in favor of an instance query answering system.
Keywords. OWL Ontology, Light-weight Description Logics, Concept Learning, Enriching Ontology.

1 Problem Description

Along with the vast use of DLs ontologies, non-standard reasoning tasks have started to emerge. One of such tasks is concept learning which is a process to find a new concept description from assertions of an ontology. The concept learning system plays an essential role in ontology enrichment as well as ontology construction. Ontology enrichment from unstructured or semi-structured data is an onerous task even for knowledge engineers. Additionally, the new added information may have diverse presentations among different engineers. As an example of concept learning, if a data set includes the assertions (John enrolled in the Semantic Web course) and (John is a Student), then a concept of "Student" can be learned by this data set which is "Who enrolled in at least one course". Therefore, this new concept definition inducted by the data will enrich the terminology of the ontology.

The current approaches of concept learning [11, 9, 20] are mostly presented for expressive DLs that are not scalable in practice. Since there are large practical ontologies that are represented by less expressive DLs such as the SNOMED CT[1], and the Gene ontology[2], it is plausible to propose a learning system for light-weight DLs that are tractable fragments of DLs in regards to standard reasoning tasks. The dedicated reasoners of light-weight DLs, such as CEL [1], Snorocket [17], and ELK [13] are very efficient for ontologies with only a TBox. These off-the-shelf reasoners do not fully support the ABox reasoning which is essential in the learning framework.

* Principal Supervisor: Professor Kewen Wang
[1] http://www.ihtsdo.org/snomed-ct/
[2] http://www.geneontology.org/

Therefore, the main research question is how to propose a learning framework to efficiently and scalably construct a concept description in light-weight description logics such as DL $\mathcal{EL}^+$ and DL-Lite. In fact, there are two main objectives for this research. The first is to design a scalable learning system which can work with real world ontologies. The second objective is to maximize the accuracy of a learned concept having incompleteness in data sets.

The remainder of this paper is organized as follows. Some preliminaries are presented in Section 2. In the next Section, the related work is investigated to find its limitations. In Section 4, the accomplished work to partially tackle the concept learning problem is presented and in Section 5, future plan followed by the evaluation of the proposed learning framework is discussed.

2 Preliminaries

An ontology in DLs consists of a *terminology* box, *TBox* $\mathcal{T}$, which represents the relationship among concepts and properties and an *assertion* box, *ABox* $\mathcal{A}$, which preserves the instances of the represented concepts and properties.

OWL EL[3], which is based on DL $\mathcal{EL}^+$ [2], is suitable for applications employing ontologies that contain very large numbers of properties and classes. In DL $\mathcal{EL}^+$, concept descriptions are inductively defined using the following constructors: $\top|\bot|\{a\}|C \sqcap D|\exists r.C$, where C and D are concept names, r is a role name, and a is an individual. An $\mathcal{EL}^+$-TBox includes general concept inclusions (GCIs) $C \sqsubseteq D$ and role inclusions (RIs) $r_1 \circ \ldots \circ r_k \sqsubseteq r$.

The DL-Lite family [5] is a family of light-weight description logics, which introduced for efficient query answering over ontologies with a large ABox, that is, the basis formalism of OWL QL[4]. Concepts and roles in DL-Lite$_R$ are constructed according to the following syntax: $B \rightarrow A|\exists R \qquad R \rightarrow P|P^-$ $C \rightarrow B|\neg C|C_1 \sqcap C_2 \qquad E \rightarrow R|\neg R$, where A denotes an atomic concept, P an atomic role, and P^- the inverse of atomic role P. B denotes a basic concept, that is either an atomic concept or a concept of the form $\exists R$. A DL-Lite$_R$ TBox is constructed by a finite set of inclusion assertions of the form $B \sqsubseteq C$ and $R \sqsubseteq E$, where B, C, R, and E are defined as above.

Note that normalized $\mathcal{EL}^+$-TBox only consists of these axioms: $A_1 \sqcap A_2 \sqsubseteq B$, $A \sqsubseteq \exists r.B$, $\exists r.A \sqsubseteq B$, $r_1 \circ \cdots \circ r_k \sqsubseteq r \in \mathcal{T}$, where $k \leq 2$, A, A_i and B are atomic concepts or $\top$. Then every existential quantifier $A \sqsubseteq \exists r.B$ in $\mathcal{EL}^+$-TBox can be replaced by these DL-Lite axioms $\{A \sqsubseteq \exists s, \exists s^- \sqsubseteq B, s \sqsubseteq r\}$.

3 Related Work

Concept learning in DLs concerns learning a general hypothesis from the given examples of a background knowledge that one wants to learn. Aiming to find a description of a *goal* concept G, there are two kinds of examples: positive

[3] `http://www.w3.org/TR/owl2-profiles/#OWL_2_EL`
[4] `http://www.w3.org/TR/owl2-profiles/#OWL_2_QL`

examples E_G^+, which are instances of G, and negative examples E_G^-, which are not. Literally, an example set of G, $\mathcal{A}'$, is a subset of ABox, $\mathcal{A}$; that is $\mathcal{A}' = \{G(a_1),\ G(a_2),\ \ldots,\ G(a_p),\ \neg G(b_1),\ \neg G(b_2),\ldots,\ \neg G(b_n)\}$, consequently $E_G^+ = \{a_1, a_2, \ldots, a_p\}$ and $E_G^- = \{b_1, b_2, \ldots, b_n\}$.

Example 1. By considering the following ABox, positive and negative examples:
$\mathcal{A} = \{$hasChild(*John, Chris*), hasChild(*Mary, Chris*), hasChild(*Joe, John*), Male(*John*), Female(*Mary*), Male(*Joe*), Male(*Chris*)$\}$
$E_G^+ = \{$*Joe, John*$\}$ $E_G^- = \{$*Mary, Chris*$\}$.
A possible answer of the concept learning problem of the goal concept "Father" is $\exists$hasChild $\sqcap$ Male.

Currently, most of the approaches to concept learning for DLs are an extension of inductive logic programming (ILP) methods. In the area of concept learning in description logics, promising research has been investigated and described in [11, 9, 20]. All of these approaches have been proposed for expressive DLs such as $\mathcal{ALC}$. One of the most significant concept learning system for DLs is DL-Learner [20] which has different heuristics to explore the search space with a built-in instance checker to employ *Close World Assumption* (CWA), that is faster than standard reasoners. However, none of these are scalable to work with real world ontologies. Nevertheless, there is little research on concept learning in DLs that transfer DL axioms to *logic program*s (LP), then apply the ILP method in order to learn a concept [10]. On the one hand, this approach is too expensive in terms of computation time. On the other hand, it is not always guaranteed that this conversion is possible. Additionally, another approach to tackle the concept learning problem in DLs is by employing a *Machine Learning* approach such as Genetic Programming [18] and kernels [8]. The experimental results of these approaches show that longer concept descriptions are generated compared with ILP based methods.

In terms of learning a concept description in less expressive DLs, research is limited. A learner for DL $\mathcal{EL}$, proposed by Lehmann and Haase [19], uses minimal trees to construct DL $\mathcal{EL}$ axioms then refines these by refinement operators. The DLs axioms were converted to trees and four different operators were defined to refine these trees. Apart from those ILP-based approaches, Rudolph [24] proposed a method based on Formal Concept Analysis (FCA) to generate a hypothesis. Further Baader et. al. [3] have used FCA to complete a knowledge base. Both of these methods used a less expressive DLs, where the former used $\mathcal{FLE}$, and the latter used a fragment of DLs which is less expressive than $\mathcal{FLE}$. These approaches demand many interactions of a knowledge engineer as an oracle of the system which is not applicable in most scenarios. In future plan, an automated system to learn new concept definitions more efficiently will be developed.

The above-mentioned approaches mostly focused on concept learning in expressive DLs, where it is not possible to have a scalable learner due to the fact that the underlying reasoners are not scalable. Therefore, a learner which produces a concept description in DL $\mathcal{EL}^+$ will be proposed, and can be employed

for DL-Lite ontologies. In the preliminary research, a learner system for DL $\mathcal{EL}^+$ using ILP-based approach and reinforcement learning technique was introduced.

4 Research Accomplished

In this section, an $\mathcal{EL}^+$ learner has been proposed since the current approaches aim to construct a concept definition in expressive DLs. However, an $\mathcal{EL}^+$ ontology necessitates the learned concepts expressed in $\mathcal{EL}^+$ only. This concept learning system is based on *inductive logic program* (ILP) techniques and finds a concept definition in $\mathcal{EL}^+$ through a *refinement operator* and *reinforcement learning* [6].

Concept Learning System using Refinement and Reinforcement: An effective tool to build the search space of concept hierarchies is requiered. According to the previous research in ILP, a refinement operator is suitable for this purpose. The proposed system benefits from the strength of the current refinement operators for $\mathcal{ALC}$ [10, 20], and a refinement operator for $\mathcal{EL}$ [19]. Downward (upward) refinement operators construct specializations (generalizations) of hypotheses [23]. The pair $\langle F, R\rangle$ is a *quasi-ordered set*, if a relation R on a set F is reflexive and transitive. If $\langle F, \sqsubseteq\rangle$ is a quasi-ordered set, a *downward refinement operator* for $\langle F, \sqsubseteq\rangle$ is a function ρ, such that $\rho(C) \subseteq \{D | D \sqsubseteq C\}$. For example, a subset of $\rho(\top)$ in the Example 1 is {Male, Female, $\exists$hasChild}, and a subset of $\rho(\exists$hasChild) is {Male $\sqcap$ $\exists$hasChild, Female $\sqcap$ $\exists$hasChild, $\exists$hasChild.Male, $\exists$hasChild.Female}. Since the refinement operator can build all possible mutations of concepts and roles, finding a correct concept description could not happen by a simple search algorithm, unless an external heuristic was employed to traverse the search space effectively. We have done some preliminary experiments in employing *reinforcement learning* (RL) technique in pruning the search space. In the proposed system, a state of a hypothesis is how correct this hypothesis is w.r.t. the given examples. This is found by the Pellet reasoner[5]. Initially, the hypothesis is the $\top$ concept. Then, an RL agent will change the hypothesis by choosing one action among those possible member of downward refinements of current hypothesis. The definition of actions is based on refinement operators that specializes the hypothesis to cover more positive examples and less negative examples. The correctness of the hypothesis, which is a score for the RL agent, will be determined by finding the instances of it. A signal is given to the RL agent according to its score to lead it to the goal state which the hypothesis is a solution of the concept learning problem. The possible actions for each state guide the RL agent to achieve the goal by this systematic reward-based approach. This approach shows promising results, however choosing an action is a non-deterministic task that causes problem where the given example sets are incomplete.

[5] `http://clarkparsia.com/pellet`

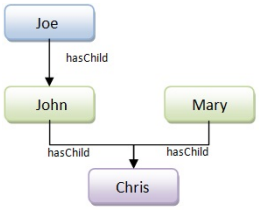
5 Future Plan

Most of the current approaches in the concept learning, including the proposed system in Section 4, use DL reasoners to accomplish instance checking task except DL-Learner which has a built-in instance checker. As a result of using OWL reasoners for the learning framework, the system becomes unscalable. Therefore, employing an efficient instance query answering (IQA) system is important for the learning framework. In this approach, query answering system is employed in order to compare certain answers of the constructed concept definition (as a query) with the given examples. A bottom-up algorithm [4] is efficiently constructed the hypothesis space, then the accuracy of any constructed concept is checked by the IQA system. An instance query (IQ) is of the form $C(x)$ with C either an $\mathcal{EL}^+$-concept or DL-Lite concept depends on learning a concept in $\mathcal{EL}^+$ or DL-Lite respectively.

Firstly, an IQA system will be developed for $\mathcal{EL}^+$ and DL-Lite queries. To achieve this, it is essential to understand how the current query answering system works efficiently. It is well-known that pure query rewriting [14] approaches are inefficient because of the exponential blow-up of the query size. Then query rewriting with auxiliary symbols [15] is introduced to include some auxiliary symbols to make the rewriting in polynomial time and this approach necessitates the saturated ABox. Our IQA is inspired by [22, 16], which complete the ABox into a canonical model $\mathcal{I}_\mathcal{K}$ of the ontology in polynomial time and independently from the input query. When $\mathcal{I}_\mathcal{K}$ can be constructed in polynomial time w.r.t. the size of the ontology, one can answer all instance queries of concepts or roles in the ontology signature efficiently. However, those auxiliary symbols cannot be the certain answer of any IQs, therefore, these unnamed individuals will be filtered from the result set.

Concept Learning System using Instance Query System: In this approach, the constructed canonical interpretation is employed as a fundamental tool to use a bottom-up algorithm in constructing a concept definition. The second research target is to construct consequence sets [12] of all positive and negative examples which are derived by IQA system. More precisely, a consequence set of an individual $a \in \text{ind}(\mathcal{A})$ is a pair $\langle rlist, clist\rangle$, where $rlist \subseteq N_R$ and $clist \subseteq N_C$ such that $\exists b \in \Delta^{\mathcal{I}_\mathcal{K}}$ with $\forall r \in rlist$, $(a,b) \in r^{\mathcal{I}_\mathcal{K}} \vee (b,a) \in r^{-\mathcal{I}_\mathcal{K}}$, and $\forall C \in clist$, $b \in C^{\mathcal{I}_\mathcal{K}}$. Then, all consequence sets of an individual a are combined as a consequence node, which is a pair $\langle rootset, conset\rangle$ such that $rootset = \{C | \mathcal{K} \models C(a)\}$ and $conset$ is the set of all consequence set of individual a. In Figure 1, the consequence nodes of the ABox instances in Example 1 are shown. Therefore, for all members of E_G^+ and E_G^-, the consequence hierarchy is constructed in order to induct a concept description. In our running example, the concept "Father" is constructed based on the common part of the consequence nodes for both *Joe* and *John* as positive examples, which in this case is "Male $\sqcap$ $\exists$hasChild", or "Male $\sqcap$ $\exists$hasChild.Male", although the second solution is subsumed by the first answer. As another example, if one wants to find a definition of the concept "Parent" with positive examples of *Joe*, *John* and *Mary*, and negative example of *Chris*, the common part of all those positive ex-

amples are ∃hasChild or ∃hasChild.Male which are correct concept descriptions for the given ontology and the example sets. Since the main interest is to find a shortest concept description, if in the first step of constructing consequence nodes a definition can not be learned, i.e. "Grandparent" in Example 1, this consequence node is extended to another step for positive examples until there is a unique common part for all consequence node of positive examples which does not overlap with any consequence node of negative examples.



Instances	Consequence Node
John	$\langle\{\text{Male}, \top\}, \{\langle\{\text{hasChild}\}, \{\text{Male}, \top\}\rangle, \langle\{\text{hasChild}^-\}, \{\text{Male}, \top\}\rangle\}\rangle$
Joe	$\langle\{\text{Male}, \top\}, \{\langle\{\text{hasChild}\}, \{\text{Male}, \top\}\rangle\}\rangle$
Mary	$\langle\{\text{Female}, \top\}, \{\langle\{\text{hasChild}\}, \{\text{Male}, \top\}\rangle\}\rangle$
Chris	$\langle\{\text{Male}, \top\}, \{\langle\{\text{hasChild}^-\}, \{\text{Female,Male}, \top\}\rangle\}\rangle$

Fig. 1. All first-step consequence nodes of the ABox instances of Example 1

6 Evaluation

The preliminary work on concept learning has been evaluated on family ontology from DL-Learner data sets[6] which is artificially constructed for test purpose and is smaller than practical ones. The proposed approach will be evaluated against current concept learning systems such as DL-Learner and YinYang[7]. There is no common benchmark for evaluating the ontology learning, although test cases have been borrowed from Machine Learning community[8] and transferred to DLs ontologies in data sets from [20]. All data sets from these concept learning systems will be used in the evaluation of the proposed approach. There are two main challenges in these benchmarks. First of all, most of the ontologies are expressed in expressive DLs, and solutions of a learning problem is not expressible by an $\mathcal{EL}$-concept description. Secondly, the second aim of this research is to have a scalable learning framework which these data sets are not applicable since the largest ontology has less than a million ABox assertions. Therefore, the LUMB benchmark[9] will be used to work on millions of ABox assertions. Some concept definitions will be removed from the TBox, then the proposed concept learning system will be applied to learn these missing concepts, and learned definitions are compared with their initial definitions. Therefore, the 'gold standard' for the

[6] http://sourceforge.net/projects/dl-learner/files/DL-Learner/
[7] http://www.di.uniba.it/∼iannone/yinyang/
[8] http://archive.ics.uci.edu/ml/
[9] http://swat.cse.lehigh.edu/projects/lubm/

learning problems are produced by querying the benchmark before the change. The completeness degree of the LUMB data sets will be tuned by another data generator [21].

As another evaluation plan, the proposed approach will be evaluated by the SNOMED CT ontology that contains more than 300K concept names, and around 60 role names in order to assess the scalability of the learning framework. However, the SNOMED CT ontology is only included a TBox which is the case for most of real world ontologies. Therefore, an ABox will be generated, for example by having different instances for all concept and role names. Then the proposed learning approach will be evaluated the same way as mentioned for the LUMB ontology by removing some definitions from the original ontology. There is also a general way of evaluating ontology learning [7], which those different metrics as quantitative evaluations will be employed in the evaluation plan.

7 Conclusion

In this paper, the concept learning problem is described to introduce its possible application in ontology enrichment. Then, two different approaches are presented for concept learning in light-weight description logics in Section 4 and Section 5. The preliminary results obtained on a small data set are encouraging which will lead to an improvement of the prototypical system to build a scalable learner. A fundamental tool to check the correctness of a learned concept definition is an instance checking system, subsequently an instance query answering system will be deployed in the proposed approach. Future work includes an implementation of the proposed approach in Section 5, as well as evaluating the scalability and efficiency of the proposed learning framework as mentioned in Section 6.

References

1. Baader, F., Lutz, C., Suntisrivaraporn, B.: CEL—A Polynomial-time Reasoner for Life Science Ontologies. In: Proceedings of the 3rd International Joint Conference on Automated Reasoning (2006)
2. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ Envelope. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. pp. 364–369 (2005)
3. Baader, F., Ganter, B., Sattler, U., Sertkaya, B.: Completing Description Logic Knowledge Bases using Formal Concept Analysis. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. pp. 230–235 (2007)
4. Baader, F., Sertkaya, B., Turhan, A.Y.: Computing the Least Common Subsumer w.r.t. a Background Terminology. Journal of Applied Logic 5(3), 392 – 420 (2007)
5. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tractable Reasoning and Efficient Query Answering in Description Logics: The *DL-Lite* Family. Journal of Automated Reasoning 39, 385–429 (2007)
6. Chitsaz, M., Wang, K., Blumenstein, M., Qi, G.: Concept Learning for $\mathcal{EL}^{++}$ by Refinement and Reinforcement. In: Proceedings of the 12th Pacific Rim International Conference on Artificial Intelligence (2012)

7. Dellschaft, K., Staab, S.: On How to Perform a Gold Standard Based Evaluation of Ontology Learning. In: Proceedings of the 5th international conference on The Semantic Web (2006)
8. Fanizzi, N., d'Amato, C.: A Declarative Kernel for $\mathcal{ALC}$ Concept Descriptions. In: The 16th International Symposium on Foundations of Intelligent Systems (2006)
9. Fanizzi, N., d'Amato, C., Esposito, F.: DL-FOIL Concept Learning in Description Logics. In: The 18th International Conference on Inductive Logic Programming (2008)
10. Fanizzi, N., Ferilli, S., Iannone, L., Palmisano, I., Semeraro, G.: Downward Refinement in the ALN Description Logic. In: The 4th International Conference on Hybrid Intelligent Systems. pp. 68–73 (2004)
11. Iannone, L., Palmisano, I., Fanizzi, N.: An Algorithm Based on Counterfactuals for Concept Learning in the Semantic Web. Applied Intelligence 26(2), 139–159 (2007)
12. Kaplunova, A., Möller, R., Wandelt, S., Wessel, M.: Towards scalable instance retrieval over ontologies. In: Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (2010)
13. Kazakov, Y., Krötzsch, M., Simancík, F.: Concurrent Classification of EL Ontologies. In: Proceedings of the 10th International Conference on The Semantic Web (2011)
14. Kikot, S., Kontchakov, R., Zakharyaschev, M.: On (in) tractability of OBDA with OWL 2 QL. In: Proceedings of the 23th International Workshop on Description Logics (2011)
15. Kikot, S., Kontchakov, R., Podolskii, V.V., Zakharyaschev, M.: Long Rewritings, Short Rewritings. In: Proceedings of the 2012 International Workshop on Description Logics (2012)
16. Kontchakov, R., Lutz, C., Toman, D., Wolter, F., Zakharyaschev, M.: The Combined Approach to Query Answering in DL-Lite. In: Proceedings of the 12th International Conference on Principles of Knowledge Representation and Reasoning (2010)
17. Lawley, M., Bousquet, C.: Fast Classification in Protégé: Snorocket as an OWL 2 EL Reasoner. In: Proceedings of the 6th Australasian Ontology Workshop (Advances in Ontologies) (2010)
18. Lehmann, J.: Hybrid Learning of Ontology Classes. In: Machine Learning and Data Mining in Pattern Recognition (2007)
19. Lehmann, J., Haase, C.: Ideal Downward Refinement in the $\mathcal{EL}$ Description Logic. In: The 20th International Conference on Inductive Logic Programming (2010)
20. Lehmann, J., Hitzler, P.: Concept Learning in Description Logics using Refinement Operators. Machine Learning 78(1-2), 203–250 (2010)
21. Lutz, C., Seylan, I., Toman, D., Wolter, F.: The Combined Approach to OBDA: Taming Role Hierarchies using Filters. In: Proceedings of the Joint Workshop on Scalable and High-Performance Semantic Web Systems (2012)
22. Lutz, C., Toman, D., Wolter, F.: Conjunctive Query Answering in the Description Logic EL using a Relational Database System. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (2009)
23. Nienhuys-Cheng, S.H., Wolf, R.d.: Foundations of Inductive Logic Programming. Springer-Verlag New York, Inc. (1997)
24. Rudolph, S.: Exploring Relational Structures via $\mathcal{FLE}$. In: Proceedings of 12th International Conference on Conceptual Structures (2004)

Semantic Interpretation of Mobile Phone Records Exploiting Background Knowledge

Zolzaya Dashdorj[1,2,3] and Luciano Serafini[3]

[1] Semantic & Knowledge Innovation Lab, Telecom Italia, Italy
[2] Department of Information Engineering and Computer Science, University of Trento, Italy
[3] Data & Knowledge Management Unit, Fondazione Bruno Kessler, Italy
dashdorj@disi.unitn.it, serafini@fbk.eu

Abstract. The increasing availability of massive mobile phone call data records (CDR) has opened new opportunities for analyzing and understanding real-life social phenomena and human dynamics. In order to better interpret this enormous amount of data it is useful to contextualize them with information about the circumstances under which they has been generated. Nowadays, linked open data initiative provide access to a huge amount of geo-time referenced knowledge about territory and events that happen in the territory. These informations can be used to characterize the aforementioned context. The aim of this Ph.D is to investigate on the intercorrelations between CDR, contexts, and human behaviors. The ultimate goal is to build a stochastic model, that can be used to predict semantic (qualitative) behavioral patterns on the basis of CDR traffic and context and identify and explain anomalous situations on the basis of deviations from standard CDR patterns.

Keywords: telecommunication dataset, stream-data, human activity, human behavior, semantics, ontology, machine learning, knowledge management

1 Relevancy

A huge amount of mobile phone data records (CDR) are generated every day in tracing users phone calls, sms, web surfing, social network interactions etc. These geo- and time-referenced data constitute an important resource of information for investigating on human behaviours. In [7, 4, 10] authors studies individual traits, human mobilities, while [11, 13, 6, 1, 8, 3] predicts communication networks and communication patterns starting from CDR. Currently, most of the analysis generate a quantitative description of human behaviors, presented via visual analytics techniques but they do not provide any insight at the qualitative/semantic level. With the term "qualitative/semantic description of human behaviour" we intend the description of human behaviors in terms of semantically rich concepts (classes and relations of an ontology) which describe, for instance, the actions performed by a person or a group of people, the events they attend, etc. From some pioneering attempts (see e.g.,[14, 1, 7, 4]]) it was clear that inferring semantically rich description of human behaviors from

pure CDR is almost impossible; It is necessary to enlarge the analysis including relevant knowledge of the context in which CDR data are generated.

Contextual information includes environmental data (e.g., weather conditions), static description of the territory (e.g., soil destination and points of interests), public and private events (e.g., concerts, sport matches, public spontaneous meeting, strikes, etc) or emergency events (e.g., accidents, traffic jams, etc), transportation schedule, energy or water consumption, etc.

This research aims at discovering the correlations between CDR stream, contexts and human behaviors, and to represent these correlations in a computational stochastic model. Using these model we can realize a set of important tasks such as: characterization of normal or exceptional events, prediction of human activities and events in certain contextual conditions (e.g., during a festival celebration that organized in the center of a city when the weather is sunny or rainy), semantic explanations to the calling or human activity distribution changes. A first attempt to create such a model has been presented in [12][4] within the Orange "Data for Development" challenge [2]. In this paper, after presenting some related work in Section 2 we describe the main thesis objective and the methodology (sections 3–6). Finally, we summarize the preliminary results and the evaluation plan and future research works (Section 8,9).

2 Related Work

The analysis of CDR by new methodologies proposed by the researchers has made great progress in the areas such as emergency response, city and transport planning, tourism and events analysis, population statistics, health improvement, economic indicators, so on [9, 2, 15, 5].

A social response to the events, in particular, behavior changes have been studied by J.P.Bagrow et al [7]. The authors explored a social response to external perturbations such emergency (bombing, plane crash, earthquake, blackout) and non emergency (festival, concert) events in order to identify real-time changes in communication and mobility patterns. The result show that under extreme conditions the level of communications radically increased right after the emergency events occur and it has long term impacts.

In [4] Calabrese et al analyzed the mobility traces of user groups with the objective of discovering standard mobility patterns associated to special events. In particular, this work analyses a set of anonymized traces of the users in Boston metropolitan area during a number of selected events that happened in the city. A result of such an analisys is that users who live close to an event are preferentially interested in that event. Similarly, Furletti et al [1] analyzes human motion associated to specific human profiles; commuter, resident, in-transit and tourist. Users are classified by a neural network, called self organizing map in one of these profiles, and the result is that the percentage of resident was compatible with the customer statistics provided by the Telecom operator. The short-ranged temporal profiles like commuter and in-transit are significantly vary

[4] http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf

and distinguishable than the larger extent profiles like resident. This analysis has been done in the city of Pisa and from the users temporal profile they identified a peak that was caused by the earthquake emergency.

Regarding the idea of collecting contextual information in the mobile phone data analysis, we propose, the similar work has been done by Phithakkitnukoon et al [14]. In this work, geographical information (Point of Interest) are collected using pYsearch (Python APIs for Y! search services) from a map. They annotated the POIs with four type of activities; eating, recreational, shopping and entertainment. The authors analyzed human activity patterns (i.e., sequence of area profiles visited by users) correlating to geographical profiles. Bayesian method is used to classify the areas into crisp distribution map of activities, which enables the activity pattern extraction of the users. The results shows that the users who share the same work profile follow the similar daily activity patterns. But not only these area profiles can explain the mobility of these users and the enlargement of activity or event taxonomy in those areas can enable the classicaions of these activity patterns.

3 Problem Statement

In this research, we are interested in analyzing the CDR in order to discover high level human behavioral patterns that can be described in qualitative/semantic terms. In other words, we generate a *semantic description of human behaviors in certain situations* when the mobile network events (phone call, sms, internet connections etc) are occurred. A semantic description of human behaviors is a representation of the behaviors of a single person, the behaviors of a group of people or of the events that happen in the human society, in terms of concepts and relations of an ontology describing human behaviors and events. An example of semantic description of human behaviors is the fact that "a person is performing some specific actions" (e.g., working, shopping, hiking), or the fact that "certain events are happening in a certain area" (e.g., a car accident, a train suddenly stops in the middle of nowhere, etc.)

To infer these types of information about human behaviors from the CDR, we need to complement these data with contextual information, which describe the context where the mobile network events are occurred. The context is a pair $\langle l, t\rangle$ where l is a location (= geographical area) and t is a time interval. For every context $\langle l, t\rangle$ from one or more knowledge repository we can extract $K_{l,t}$ which is a knowledge base describing this context. $K_{l,t} \equiv O_l \cup E_{l,t}$ i.e., it is the union of the objects which are nearby the location l (point of interests), and the events which take place nearby l at time around t. Every element of O_l is a pair $\langle poi, w\rangle$ where $w \in [0, 1]$ is a weight that expresses the closeness of poi to l. Every element of $E_{l,t}$ is a pair $\langle e, w\rangle$ with $w \in [0, 1]$, expresses how close to l and t is the event. Examples of POIs are buildings, roads, natural points, shops, etc., examples of events are weather phenomena (rain, snow, etc.) or social events, like concerts, strikes, traffic jams, etc. For every context, from the knowledge repository about the context, we can derive $A_{l,t}$ which are the most probable human activities in the context $\langle l, t\rangle$. Every element of $A_{l,t}$ is a pair $\langle a, w\rangle$ where

$w \in [0, 1]$ is a weight that expresses the likelihood of a parson performing the activity a in the context $\langle l, t \rangle$. Examples of the activities are working, studying, shopping, attending in a concert, travelling by car, etc

Making use of the knowledge repository about contexts, we enrich the CDR with human activities and events. Contextually enriched CDR can be exploited to analyse and evaluate call patterns associated to human activities and events. A call pattern is a quantitative model that describes the "normal" behaviour of a communication in a certain class of contexts. An example of a call pattern is the function that associates the number of calls done by people before, during, and after a particular event (e.g., a football match, or a concert). Another example of call pattern (also called interaction pattern) describe the number of calls between a pair of locations before, during one particular time of the working day. A third example of call pattern (also called mobility pattern) describes the number of user displacement from one location to another. Call patterns describes the normal behavior patterns. Comparing call pattern with the actual calls allows us to identify divergent and exceptional behaviors and context allows us to characterize such behaviors (e.g., offering explanations for these behaviors) and the prediction of similar patterns.

4 Research Questions

A more concrete formulation of the research questions are presented below. The questions are posed to discover the correlations between CDR stream contexts and human behaviors and to represent these correlations in a computational model, that can be queried to obtain the following informations:

RQ1 *What are the correlations between contexts and human behaviors? What are the most probable action that a person is doing in a given context?*

RQ2 *What are the correlations between contexts and CDR? What is the normal call frequency pattern in a specific context (where a context can be, a type of area, an event, a time of the day etc..)?*

RQ3 *What are the correlations between call frequency and human behaviors? What is the call frequency pattern of people while performing certain actions?*

To achieve the answers for these core questions, we need to engage the preliminary challenges in linking the data coming from different datasources and real-time knowledge reasoning and pattern recognition in streaming data from the view of computational and conceptual perspectives.

5 Hypotheses

The main hypothesis to this Ph.D research work is to characterize a human behavior that represented in the form of call frequency pattern as it is connected to the semantics of human activities or events in a certain context. The multi-classification of these human behaviors recognizes the situation changes that engaged in the context (e.g., when the weather is sunny or rainy in a certain location and time). This improves the prediction task of human behaviors.

6 Methodology

We organize the work for the semantic interpretation of human behavior in mobility based on the merge of mobile network data stream and the geo and time referred available background knowledge in the following phases:

1. The starting point is a characterization of the territory with the human activities and events that can be performed in any location of an area and at any time of a day. This phase is intended to answer the RQ1 described in Section 4.
2. On the basis of the correlation obtained in phase one, we annotate mobile phone network events with the human activities or events and then extract the semantic behaviors determining the typical calling patterns that associated to various typical activities or events. This answers the RQ2,3.
3. The result of the second phase can be used to explain the possible or anomalous actions and situations in real-time, when the call activity sensibly deviates from the standard call activity associated to a known context.

6.1 Geographical Area characterization

We characterize the territory with contextual information in which mobile network events are occurred, in order to model the relation between human activities and contexts. All the possible contextual information can be retrieved through the employment of online and offline techniques in information retrieval and text mining, probability inference and those information can be scored in order to determine the importance. The example of contextual information include mobile cell covereage map, POIs distribution, social event distribution and domain statistical data about demography, ethnography, energy or water consumption, so on. The allocation of contextual information to each context enables analysis and identification of the possible human activities or events associated with a likelihood which expresses a probability of the activity that could be performed by the users. For example, an area which contains mostly about highway and if there is an accident on the highway, the probable event is a traffic jam while people are performing an action, "travelling by car". For modelling the relation between contexts and human activities, we propose two steps; (1) an ontological model that describes the concepts and the relations between the ontologies of human activities and knowledge about contexts, under the expert knowledge derived from surveys, crowdsourcing and domain experts (2) a stochastic behavior prediction model that predicts the possible top-k activities or events associated with a likelihood that could be performed in each context. The approach quantifies the correlation between contexts and human activities through uncertain probabilistic modeling techniques such as Probabilistic Boolean Networks, Markov Logic Networks, Bayesian Networks. The approach can enable a further consideration of the OWL language extensions with the probability of the activities or events in order to do reasoning with OWL for the prediction.

6.2 Extraction of Semantic Behavioral Patterns

On the basis of the previous association, context → human activities, in this step, we propose a semantic behavioral analysis model that annotates the CDR to the most probable actions/events that happening when the phone calls are localized in the territory. By adapting the state of the art techniques of behavioral call frequency pattern extraction in the area of mobile phone data analysis, we extract a standard type of call frequency patterns about human mobility, communication and interaction patterns that annotated with the human activities/events. The extracted semantic behavioral patterns are classified into certain types based on the similarity metrics of the call frequency patterns which characterized with the contextual knowledge, making use of classification techniques such as Logistic regression, Naive bayes, Perceptron, SVM, and novel classifier fusion methods so on. The classifcation results are stored into a behavioral decision tree repository in each area. This explains the correlation between CDR and contexts as well as human activities or events. Example of semantic behavioral pattern is attending in a concert can be vary depends on the weather condition, geographical location and the events that could be occurred at the same time, so on.

6.3 Forecast of the CDR stream in Real-time

Exploiting those models in real-time, we will forecast the CDR stream in the given territory to explain the possible or anomalous actions and situation changes in real-time. We adapt algoritms which operate in online and incremental fusion such as streaming linked data framework, C-SPARQL that will propose online techniques for annotating the semantic labels of contextual information that described in the form of Linked Open Data with the call frequency patterns in CDR stream. The CDR stream can be transformed into RDF stream and the reasoning of the stream can be done. The observed call frequency patterns are analysed and reasoned comparing to the classification of semantic behavioral patterns that stored in the behavioral desicion tree repository in each area. This enables identification and prediction of standard or anomalous type of behaviors in real-time offering semantic explanations to the CDR stream. The new discovered behaviors can be learned to the behavioral desicion tree repositories.

7 Reflections

This Ph.D research work is intended to develop a novel model that deepens the analysis of the CDR through machine learning approaches and logical semantics considering a wide range of contextual features in each context where mobile network events occur. This could provide an extensive overview to the CDR and that could be interpreted in qualitative terms.

8 Evaluation Plan

An evaluation will be divided according to each phase of the methodology that described in Section 6:

Phase1 At the first phase, we evaluate the model that quantifies and qualifies the correlation between human acivities and contexts based on user data (ground truth) we have collected through a web or mobile phone application, about daily activities that performed in various areas of the territory and in different times of a day. We choose several cities as use case in order to do comparative analysis. The preliminary evaluation has been done. In this evaluation, the model charaterizes every context of the mobile network events in the Trento city, Italy with the possible human activities that extracted from a geo-referenced datasource, OpenstreetMap. By collecting user-feedback, we obtained 70.89% of overal accuracy, and 61.95% of overal accuracy among the top-5 activities.

Phase2 At this phase, the evaluation is concentrated in the correlation between CDR and contexts as well as human activities. We use a sample CDR dataset (training, test) that covers particular events and festivals, concerts, etc that selected for the evaluation. In training dataset, the semantic behavioral patterns are extracted and classified into certain types.. The test dataset is used to measure the accuracy of the performance of the classfication model for predicting semantic behavioral patterns considering the call frequency patterns associated with similar events, festivals and concerts, etc. We compare the results in different cities.

Phase3 We concentrate in the real-time analysis of the CDR stream to evaluate the performance of identification and prediction of possible or anomalous actions and situations in real-time. We use the training dataset which used in Phase2 in order to use the semantic behavioral patterns for identification and prediction of actions in the CDR stream. We evaluate the resuls with a help of domain experts.

9 Preliminary Results and Conclusion

In this paper, we presented the Ph.D work aimed at understanding the correlations between CDR, human behaviors, and contexts in a computational model. To understand these correlations from quantitative data (CDR), we complement contextual information in order to describe the context where a phone call is done. Our methodology which addresses these problems is divided into three core phases 1) geographical area characterization 2) extraction of semantic behavioral patterns 3) forecast of the CDR stream in real-time. At the first phase of the evaluation, we obtained 70.89% of overal accuracy, and 61.95% of overal accuracy among the top-5 activities.

Next step of this Ph.D work is to enrich the geographical area characterization making use of various type of geo/time-referenced contextual information available on the web sites such as environmental data about weather condition, and public and private events about festivals and concerts and emergency events about accident or strike and some other domain statistical data about energy consumption, so on. We organize a wide range of evaluation for this model, involving as many as participants who can share their daily activities. On the

basis of this model, we will work on the following phases; extraction of semantic behavioral patterns and forecast of the CDR stream in real-time that will allow us to determine semantically rich heterogeneous (normal or anomalous) human behaviors in real-time.

References

1. B.Furletti, L.Gabrielli, C.Renso, and S.Rinzivillo. Identifying users profiles from mobile calls habits. In *the Proc. of the ACM SIGKDD Int.Workshop on Urban Computing*, UrbComp '12, pages 17–24. ACM, 2012.
2. Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.
3. Nathan Eagle and Alex (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006.
4. F.Calabrese, F.C.Pereira, G.Di Lorenzo, L.Liu, and C.Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *the Proc. of the 8th Intl.Conf. on Pervasive Computing*, Pervasive'10, pages 22–37, 2010.
5. Ferrari.L, Berlingerio.M, Calabrese.F, and Curtis-Davidson.B. Measuring public-transport accessibility using pervasive mobility data. *IEEE Pervasive Computing*, 12(1):26–33, 2013.
6. J.Candia, M.C.González, P.Wang, T.Schoenharl, G.Madey, and A.Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, June 2008.
7. J.P.Bagrow, D.Wang, and A.Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.
8. R. Lambiotte, V. Blondel, C. Dekerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Vandooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, September 2008.
9. K Laurila.J, Gatica-Perez.D, Aad.I, Blom.J, Bornet.O, T. Do, Dousse.O, Eberle.J, and Miettinen.M. The mobile data challenge: Big data for mobile computing research. Newcastle, UK, 2012.
10. M.C.Gonzalez, C.A.Hidalgo, and A.Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.
11. J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks, 2006.
12. P.Paraskevopoulos, T.Dinh, Z.Dashdorj, T.Palpanas, and L.Serafini. Identification and characterization of human behavior patterns from mobile phone data, 2013.
13. Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.
14. S.Phithakkitnukoon, T.Horanont, G.Di Lorenzo, R.Shibasaki, and C.Ratti. Activity-aware map: identifying human daily activity pattern using mobile phone data. In *the Proc. of the 1st Intl. Conf. Human Behavior Understanding*, pages 14–25, 2010.
15. A. Wesolowski, N. Eagle, A.J. Tatem, D.L. Smith, A.M. Noor, R.W. Snow, and C.O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–70, 2012.

Ontology-based top-k query answering over massive, heterogeneous, and dynamic data*

Daniele Dell'Aglio

Dipartimento di Elettronica, Informazione e Bioingegneria – Politecnico of Milano,
P.za L. Da Vinci, 32. I-20133 Milano - Italy
daniele.dellaglio@polimi.it

Abstract. A relevant kind of task, which is getting more and more attention in the recent years, is the selection of the most relevant elements in a number of data collections, e.g., the opinion leaders given a set of topics, the best hotel offers, and the best cities to live in. At the moment the problem is addressed through ad-hoc solutions tailored on the target scenario, setting up infrastructures able to manage the expected data loads. Even if these solutions work, they could start to have problems when the data volume, velocity and variety increase. In my research activity I will study the problem of computing the top k relevant items given a collection of data sets with both streaming and static data, an ontology describing them, and a set of top-k queries where each scoring function describes the relevance as a combination of several criteria.

1 Introduction

Relevancy. Big Data are characterized by the so-called three Vs [1]: volume (high amount of data), velocity (highly dynamic in data) and variety (data with structural and semantic heterogeneity). When those data are analysed and queried, it often happens that there is a huge number of answers, but only a small part of them is relevant (on the basis of some criteria). Let's consider, for example, Expedia[1]: it processes travel solutions and it provides as search result the top k results (i.e., the first result page) ordered by a set of user-provided criteria. In most of the cases, users find the solution they are looking for in the first page, without moving to the next ones. The input data used by Expedia can be described through the Big Data dimensions[2]. Variety is given by the fact that data are related to different domains (flight companies, hotels, and car rental services), are gathered by several sources and have to be integrated. Velocity is a critical dimension for the Expedia service: availability and price of flight tickets and hotel rooms are dynamic and the quality of the returned results is strictly

* This research is developed under the supervision of Professor Emanuele Della Valle.

[1] Cf. http://www.expedia.com

[2] Even if the whole data computation process is not made by Expedia and there are intermediate data providers (e.g., Amadeus), I consider this process as executed by a black box system to highlight the features of the data.

dependent on it. Finally, data about about flights, hotels and car rental have relevant volumes.

Even if Expedia does not have Big Data problems (its infrastructure is able to cope with the data it processes), a system like Expedia that pushes further the aforementioned input data dimensions (e.g., use more domains and more dynamic data) could have problems in maintaining the quality of the supplied services. The main goal of my research is the development of methodologies and algorithms for computing top k results (ordered by a given criteria) in a Big Data scenario.

In the last decade, the research activities in this domain have addressed different sub-problems: the finding the most relevant elements can be expressed through *top-k queries*, i.e., queries that asks for the top k tuples from a dataset, given an order expressed through a scoring function [2]; the velocity is addressed by *stream computation* and *on-line streaming algorithms* to process data in real time [3]; data variety and data access can be addressed through *ontologies* to obtain an holistic view on heterogeneous data sets, exploiting the Ontology Based Data Access (OBDA) approach [4].

How to combine these methods and techniques is an open research issue: RDF stream engines [5], top-k ontological query answering [6] and top-k computation over data streams [7] are examples of novel research trends that are gathering more and more attention in the recent years. In my activity I will study how those elements can be combined in order to efficiently perform data analyses in a Big Data context.

Problem Statement. Figure 1 depicts the framework I will consider to probe the problem presented above. Data are gathered from collections of data sets $\mathbb{D}$ and data streams $\mathbb{S}$. The data model is described by an ontology $\mathcal{M}$, and the information needs are defined through a set $\mathbb{Q}$ of continuous queries, containing a subset $\mathbb{K}$ of top-k continuous queries. The problem I investigate in my activity is how to optimize the query answering in this setting.

It is worth to note that the existence of a set $\mathbb{Q}$ of queries is not a stretch: in stream processing applications is common to develop network of queries [8], where each query produces streams and consumes outputs of other queries.

The remaining of the paper is structured in the following way: Section 2 describes the related works; Section 3 presents the research questions and the relative hypotheses that will drive my research activity. Section 4 describe the approach I follow to test the hypotheses and Section 5 ends with some final considerations.

2 Related work

The raising of data stream sources introduced new problems about how to manage, process and query infinite sequences of data with high frequency rate. Two proposed approach are the Data Stream Management Systems (DSMSs) and Complex Event Processors (CEPs) [3]: the firsts transform data streams in timestamped relations (usually through the *window* operator) to be processed with

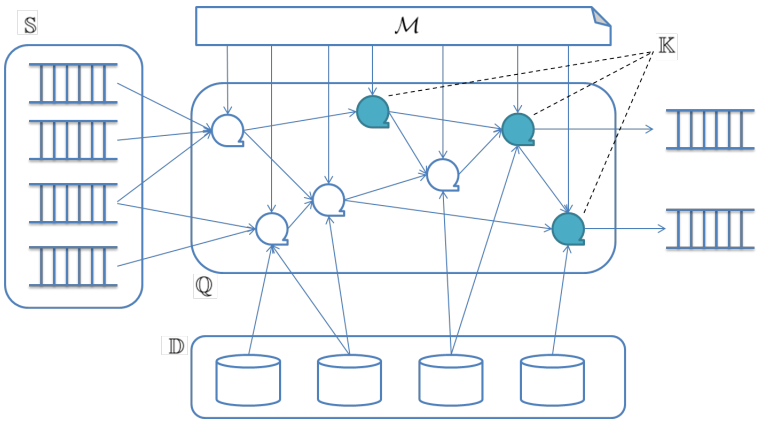


Fig. 1. Elements involved in the problem.

well known techniques such as algebras [8]; the seconds look for patterns in the streams to identify when complex events occur [9]. Recently, those paradigms have been studied by the Semantic Web community, that come out with different relevant results. On the one hand, the DSMS model inspired the design and the development of C-SPARQL [10], SPARQL$_{stream}$ [11] and CQELS [12]. On the other hand, the CEP model inspired the EP-SPARQL [13] system. RDF stream engines are at the basis of my research: they are the first step towards data streams and ontologies[3].

A parallel topic that joins Semantic Web and Stream Processing is the execution of reasoning tasks over data streams. The work in [14] proposes a method to solve reasoning tasks over an ontology stream (i.e., a sequence of timestamped ontologies). The work in [15] focuses explicitly on the ontological query answering in RDF stream processors: it proposes an algorithm, inspired to DRed [16], to incrementally maintain the ontological entailment of the content of a window. The approach is data-driven, and the entailment is updated when the window slides.

The top-k query answering problem has been widely studied in the Data Base Management System area [2]: the general idea is to extend the algebras introducing top-k related operators as first-class citizens, and to provide physical operators able to determine the top k tuples, ordered by a given criteria, without scanning the whole dataset.

Top-k query over data streams and top-k ontological query answering are research trends in an initial stage. [7] is one of the first works focusing on how to maintain a top-k answer over a stream. Regarding the top-k ontological query answering, SPARQL-RANK [17] proposes an extension of SPARQL to optimize

[3] In the following, with RDF stream engines I will indicate the RDF stream engines following the DSMS paradigm.

the top-k query answering; anyway, the approach works only under the RDF entailment regime. SoftFacts [18] is a top-k retrieval system that uses an ontology layer to manage the conceptual model (using OWL-QL as ontological language), and relational database systems to store and query the data.

3 Assumptions, research questions and hypotheses

The operational semantics of RDF stream engines such as C-SPARQL, CQELS and SPARQL$_{stream}$) can be described by the model proposed by Stream and CQL [8], and depicted in Figure 2.

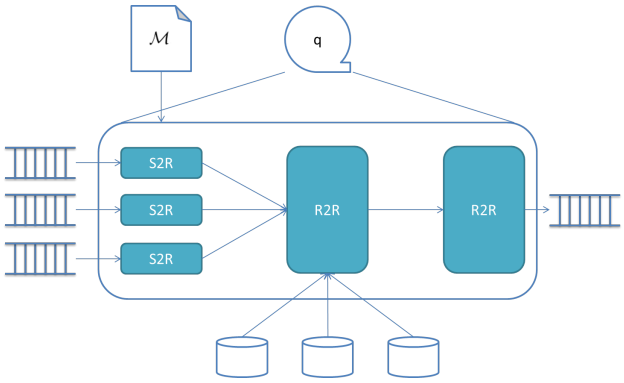


Fig. 2. General model of a continuous SPARQL query

The query answering process can be represented in three logical steps. First, the input RDF streams are transformed in sets of mappings (using SPARQL algebra terminology) through the S2R operators[4], usually sliding windows. The resulting sets of mappings and data from static data sets are transformed in a new set of mappings through a R2R operator, (the boolean expression part of the query, compliant with SPARQL 1.1/1.0). Finally a R2S operator converts the mappings in the output stream.

The reasoning technique in [15] works on the window operator: it uses the ontology $\mathcal{M}$, the window content and the static data to compute the materialization. The latter is then used as input by the R2R operator. This materialization method works under some assumptions:

- TBox assertions are not in the input stream;
- the input of the query is one window over one stream (and optionally static data sets);

[4] I maintain the operators' names as defined by CQL: S2R (stream-to-relation), R2R (relation-to-relation) and R2S (relation-to-stream).

– the same statement cannot be in both the input stream and the static data.

At the beginning of my research these assumptions hold, and I plan to investigate if they can be relaxed (in particular the second and the third ones).

Additionally, I make an assumption on the static data sets: they are available in the first or the secondary memory of the machine that execute the query. Top k algorithms need fast accesses to the data sources, and the problem of designing algorithms to execute top k queries over remote repositories (exposed via SPARQL endpoints) is out of scope in my activity.

Research questions. Starting from the problem statement and exploiting the RDF stream engine paradigm presented above, I define the following research questions:

Q.1 In RDF stream processors the query are registered before the arrival of the data, and they provide sequence of time ordered results depending on the content of the windows. How can these facts be exploited to: 1) design more efficient algorithms and 2) improve the expressiveness of the ontological language used to define the ontology $\mathcal{M}$?

Q.2 The queries in $\mathbb{Q}$ are translated in a set of logical plans (algebras) and then in physical plans to be executed. How the topology of the query network, the model $\mathcal{M}$, and the scoring functions defined by the top-k queries in $\mathbb{K}$ can be used to optimize the plans?

The two questions aim at probing the optimization of the query answering process by two different points of view. Question Q.1 focuses on how stream processing engines manages the queries: the queries in $\mathbb{Q}$ are registered in the system before the data arrive, and each query is evaluated multiple times on different portions of the input streams. Question Q.2 takes into account the query plans optimization, and in particular how the ontology $\mathcal{M}$ and the scoring functions in top-k queries $\mathbb{K}$ can be used to improve the logical plans and the physical plans of the queries.

Hypotheses. In the attempt to investigate the answers for these research questions, I formulated a set of hypotheses that will lead my activity. Regarding the question Q.1, the hypotheses are:

H.1.1 The available stream reasoning techniques can be extended to work with queries with multiple windows.

H.1.2 The available stream reasoning techniques can be optimized when there are multiple queries.

H.1.3 Due to the fact that both the conceptual model and the query are fixed, it is possible to improve the expressiveness of the ontological language used to define $\mathcal{M}$, maintaining the query answering problem over data streams treatable.

H.1.4 It is possible to move from a purely data-driven approach (i.e., materialization) to a hybrid query-driven and data-driven approach (i.e., query rewriting and materialization) to improve the memory consumption and the response time.

The four hypotheses are related to the query answering process and how it can be improved. Stream reasoning technique in [15] assumes the existence of one window (and optionally a static data source): with hypothesis H.1.1 I want to verify if and under which conditions it is possible to relax this constraint. The hypothesis H.1.2 test if, given the set $\mathbb{Q}$ of queries, it is possible to make synergies and do not maintain $|\mathbb{Q}|$ materializations separately. Hypothesis H.1.3 is related to the expressiveness of the ontological language: the technique presented in [15] works on RDFS+, but it could be possible to extend it to support more expressive languages. Finally, through the hypothesis H.1.4, I want to test if in the stream context it is possible to exploit the query registration process to rewrite the query using $\mathcal{M}$, reducing the memory consumption of the materialization and the time required to maintain it.

In parallel, I aim to probe the question Q.2 through the study of the following hypotheses:

H.2.1 Exploiting the network of queries, it is possible to optimize the query plans in RDF stream engines.
H.2.2 The presence of $\mathcal{M}$ and $\mathbb{K}$ allows to optimize the query plans at logical level.
H.2.3 It is possible to exploit $\mathcal{M}$ and $\mathbb{K}$ to design physical operators that perform faster.

In this set of hypotheses I focus on the optimization of the query plans. In stream processing engines, such as Aurora [19], groups of query plans are optimized; in hypothesis H.2.1 I want to test if optimization techniques for stream processing engines can be applied and extended to RDF stream engines. Hypothesis H.2.2 and H.2.3 aim at testing that the ontology $\mathcal{M}$ and the top-k queries $\mathbb{K}$ enable optimizations in the query plans (respectively at logical and physical level).

Reflections. The problem I am going to investigate is the optimization of continuous ontology-based top-k query answering. As explained above, even if sub-problems are addressed, at the best of my knowledge there are no results that address this problem. I think that the main reason is that methods and instruments at the basis of this activity have lacked until some time ago. For example, RDF stream processing engine and stream reasoning topics are novel and open research trends, and the research groups of these fields put the most of the effort in principles and foundation definitions. I believe that these results compose a solid basis to build my research activity in the next years.

4 Research Plan

Approach. My research starts from a deep state-of-the-art analysis on data management and description logic fields; some relevant results are presented in Section 2. In parallel, there will be the identification of a set of use cases, to determine a set of real problems from which elicit the requirements that will lead my activities. At the moment I am considering a social listening scenario:

the idea is to analyse social networks data and mobile phone records to extract knowledge.

The next step of my activity is the design of the evaluation framework and the evaluation metrics. Even if some hypotheses will be investigated through a theoretical approach, others will require an empirical approach. As consequence, it is important to define this tool from the beginning: it allows to measure the progresses of the activity and to evaluate the experiments.

As result of the state of the art analysis, the requirements elicitation and the evaluation framework definition, there will be an environment to support the core activity of my research, the tests of the hypotheses. I will follow a three-step plan. In the first step, I will focus on the hypotheses related to question Q.1, and consequently on continuous ontological query answering over multiple streams. In this phase the top-k element is missing: at the moment the research on top-k query answering is more mature than the one on inference over data streams, so I believe it is necessary to work on the latter before studying the synergies between them. In the second step, I will target Q.2, through the test of the relative hypotheses. Finally, in the third step I will bring together the results obtained. The output of each step will be a set of approaches, supported by prototypes to prove their feasibility and to evaluate them.

Evaluation Plan As explained in the previous section, an evaluation framework is required since the first steps of my activity. The metrics I aim at defining are related to the following dimensions: response time (the time required to compute the answers of the query), memory consumption, and correctness of the results.

In the previous months I started to work on the design of the evaluation framework. The starting point was the analysis of benchmarks for RDF stream processors. At the moment, at the best of my knowledge, there are two available benchmarks: SRBench [20] and LSBench [21]. The two works address different features of the RDF stream processors, such as the degree of SPARQL 1.1 adoption, the time performance and the throughput. A feature that has not been investigated yet is the correctness of the results provided RDF stream engines. I started to work on this aspect, and [22] reports some initial considerations. The most relevant one is that the available operational semantics of the RDF stream engines are not enough to model the different behaviours shown by the systems. As consequence, I am contributing in defining a framework to validate the results provided by the system.

5 Conclusion

In the next years, the Big Data market will grow [23], and, as consequence, approaches and methodologies to manage, process and extract knowledge will become more and more important. The problem I am going to investigate is the optimization of the top k elements retrieval task in a context characterized by heterogeneous and massive data streams. I presented the research questions and the hypotheses that will lead the activity, and the plan I will follow to address them.

References

1. Laney, D.: 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group (February 2001)
2. Ilyas, I.F., Beskales, G., Soliman, M.A.: A survey of top-k query processing techniques in relational database systems. ACM Comput. Surv. **40**(4) (2008)
3. Cugola, G., Margara, A.: Processing flows of information: From data stream to complex event processing. ACM Comput. Surv. **44**(3) (2012) 15
4. Lenzerini, M.: Data integration: A theoretical perspective. In: PODS. (2002) 233–246
5. Della Valle, E., Ceri, S., van Harmelen, F., Fensel, D.: It's a streaming world! reasoning upon rapidly changing information. IEEE Intelligent Systems **24**(6) (2009) 83–89
6. Schlobach, S.: Top-k reasoning for the semantic web. In: ISWC. (2011) 55–59
7. Mouratidis, K., Bakiras, S., Papadias, D.: Continuous monitoring of top-k queries over sliding windows. In: SIGMOD Conference. (2006) 635–646
8. Arasu, A., Babu, S., Widom, J.: The CQL continuous query language: semantic foundations and query execution. VLDB J. **15**(2) (2006) 121–142
9. Luckham, D.C.: The power of events - an introduction to complex event processing in distributed enterprise systems. ACM (2005)
10. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: C-SPARQL: A continuous query language for RDF data streams. IJSC **4**(1) (2010) 3–25
11. Calbimonte, J.P., Corcho, O., Gray, A.J.G.: Enabling ontology-based access to streaming data sources. In: ISWC. (2010) 96–111
12. Le-Phuoc, D., Dao-Tran, M., Xavier Parreira, J., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. In: ISWC. (2011) 370–388
13. Anicic, D., Fodor, P., Rudolph, S., Stojanovic, N.: EP-SPARQL: a unified language for event processing and stream reasoning. In: WWW. (2011) 635–644
14. Ren, Y., Pan, J.Z.: Optimising ontology stream reasoning with truth maintenance system. In: CIKM '11
15. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Incremental reasoning on streams and rich background knowledge. In: ESWC (1). (2010) 1–15
16. Volz, R., Staab, S., Motik, B.: Incrementally maintaining materializations of ontologies stored in logic databases. J. Data Semantics **2** (2005) 1–34
17. Magliacane, S., Bozzon, A., Della Valle, E.: Efficient execution of top-k sparql queries. In: ISWC. (2012) 344–360
18. Straccia, U.: Softfacts: A top-k retrieval engine for ontology mediated access to relational databases. In: SMC. (2010) 4115–4122
19. Abadi, D.J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., Zdonik, S.B.: Aurora: a new model and architecture for data stream management. VLDB J. **12**(2) (2003) 120–139
20. Zhang, Y., Duc, P., Corcho, O., Calbimonte, J.P.: SRBench: A Streaming RDF/SPARQL Benchmark. In: ISWC. (2012) 641–657
21. Le-Phuoc, D., Dao-Tran, M., Pham, M.D., Boncz, P., Eiter, T., Fink, M.: Linked stream data processing engines: Facts and figures. In: ISWC. (2012) 300–312
22. Dell'Aglio, D., Balduini, M., Della Valle, E.: On the need to include functional testing in rdf stream engine benchmarks. In: BeRSys 2013. (2013)
23. Gens, F.: IDC Predictions 2012: Competing for 2020 (2012)

ISWC 2013 Doctoral Consortium 'Ontology Evolution for End-User Communities'

Peter J. Goodall and Peter Eklund

Centre for Digital Ecosystems
University of Wollongong,
peterg@acm.org, peklund@uow.edu.au

Abstract. This project supports application for a *practice-based*[4] PhD by Peter Goodall. The project will produce a system architecture and proof-of-concept laboratory implementation to model, instrument, prototype, and evaluate the effectiveness of several alternate system designs which are intended to enable small end-user communities to evolve specialized ontologies and annotations for entities important to them. Depending on available resources, the laboratory may also be used to study bridging-subset ontologies for interchange, federation and seeding of community ontologies. There will be a discussion of the strengths and weaknesses of various approaches tried by others, and a reflection on the usefulness of the system resulting from innovative work of this project.

Keywords: information systems, ontology, taxonomy, emergent systems, digital ecosystems, curation

1 Problem Description

Each living human community, small or large, cultural, scientific or commercial has its characteristic evolving shared nomenclature which enables discussion and action involving entities and activities important to that community. Further, communities that interact or exchange information with each-other require intersecting terminologies.

Ontology is the concept which encompasses terminology with its meanings and mechanisms. Ontology has philosophical and computational streams of interest to this project: *Philosophical Ontology* has been described as 'that branch of metaphysics concerned with the nature or essence of being or existence' [17]. Its modern child - *Computational Ontology* can be described as 'a conceptualization of some universe of discourse, embodied as a declarative formal abstraction'[14].

This project is motivated by observation of a contradiction between the curatorial and cultural perspectives of an Australian Research Council Linkage Grant (now inactive) which I project-managed - known as 'The Virtual Museum of the Pacific'[10,9] (no longer active).

The Cultural Collections section of our project partner - The Australian Museum, has over many decades iterated through developing or adopting controlled

vocabularies for cataloging their Pacific Collection. They currently use a concise in-house taxonomy for very practical reasons.

While preparing for the launch of the project web-site, we spent time working with representatives of some Pacific communities. They were concerned that the collection taxonomy had little relationship with their own way of describing their artifacts within the collection.

Both the cultural and curatorial ontologies struggled to be effective in the context of the collection. Both the curatorial staff and cultural owners were experts in the conceptualization of their particular perspectives of the collection, and both need aid in ongoing development of terms to suit their both separate and overlapping needs.

My project goal is to research and develop systems for exploring ontologies computationally emergent from technically-informal community annotation.

2 Relevancy

This project has both social and commercial relevancy:

Social - many communities of commitment or interest [21,11] not equipped or resourced to create formal ontologies for curating their physical and electronic artifacts, even though their members are the primary domain experts of those communities.

Commercial - because nearly every modern large-scale electronic market or business, has a shortage of technical experts to formally classify objects and data for cataloging and recommendation.

Both of these application areas could greatly benefit from further development of categorization generated from community or customer based annotation.

3 Related Work

Eleanor Rosch and others, in a number of seminal papers [18,19] developed the notions of Basic Categories and Prototype Theory, prosecuting the idea that there are a number of fundamental levels of category that have a degree of coherence amongst end-users. This experimental work inspires some confidence in community labeling of objects being used to produce derivable category structure. More recently work on the Rational Model of Categorization and the Feature Induction Model examined by Sanborn [20] show promise for inferring categories from user tags.

The work of Cimiano and colleagues [6,7] provides a theoretical and practical resource for generating Formal Concept Lattices for adaption to work done in our *Virtual Museum of the Pacific* project and its Formal Concept Analysis-based navigation and tagging [8].

Mika [15] provides an extended ontology tripartite hypergraph model, splitting the hypergraph into three bipartite graphs associating actors with concepts, concepts with instances and actors with instances. These bipartite graphs are

then folded using matrix operations to create various useful affiliation networks, whose properties are used to develop lightweight ontologies demonstrated via a number of case-studies.

A data-model for capture, archiving and processing of user tagging events requires some thought. A user labels an object while engaged in some cognitive context, the person chooses a symbol relevant to that context as a reminder of the concept from that context. Some investigation of processes and schools of Semiotics [5,16] leads to a semiotic triad augmented with a time-stamp and some flexible metadata.

The time-trace of these tags bears an interesting affinity with Dynamic Topic Modeling and Topic Hierarchies [2,3,12] - an approach for incremental discovery of hierarchic categories which appears worth investigating.

As noted in the Problem Description of this overview, development of category systems by domain experts who are not ontologists is discouraged by the complexity of traditional approaches. This complexity is reflected in the software tools readily available. Protege [13] is an example of a quality open-source ontology development system, many others are available in various states of development usage and repair [1] .

4 Research Questions

1. How to model simple hierarchies of community generated categories?
2. What is the relationship between community and classification?
3. Are time-based data-sets available which demonstrate category drift within a community over time?
4. How much information external to the tagging discourse is required for an ontology to emerge?
5. Is emergence as a computational model valid?
6. How to represent collections of tagging events in a way that is simple and effective, yet consistent with *relevant* semiotic theory?
7. How to model requirements for selecting suitable experimental datasets and feeds
8. How to access and transform datasets/feeds for ingestion into the system, and select appropriate tool-sets for this?
9. How to represent and derive ontologies from tagging event streams?
10. Which formal models of tags and ontologies are suitable for incremental and bulk computation?
11. Define a system architecture for processing, interaction, and visualization.

5 Hypotheses

These hypotheses are still very much under development.

1. Basic-level ontologies can be computationally emergent from a suitably annotated tagging event stream. The emergent ontologies will be judged useful by the annotating community.

2. Ontologists or curators working in the domain of a community's emergent ontologies (as derived above) will judge those ontologies to be useful to their work.
3. A proof of concept technical work-bench/tool-chain can be constructed from Free and Open Source Software (FOSS) and reasonably generic cloud computing resources, which can be used to effectively experiment with extraction of latent categories in tagging discourses.
4. Cohesiveness of a subject domain's community of interest corresponds to the coherence of the vocabulary used when tagging objects from that domain.
5. A cohesive community of interest tagging their domain of interest, will produce a stream of tagging events whose latent categorizations will converge into a set of basic and superordinate categories [19], which are recognized as useful by that community.
6. A tagging event data-set with a long-enough time-base will demonstrate temporal drift of its latent categories.
7. Scalable algorithms can be found to derive basic and superordinate categories incrementally from suitable annotation event streams.

6 Approach

1. Literature Survey - Research literature to understand
   - the fundamental relationships between semiotics, classification, community, crowd-sourcing, folksonomy.
   - computational topics - such as Formal Concept Analysis, ontology extraction from text, functional programming, analysis packages such as R, Incanter and Pandas.
2. Characterize and discover suitable data-sets. It is important to find, where possible, suitable existing data-sets for testing hypotheses before committing to the expense and difficulty of user-testing.
   - There are a number of ad-hoc web-sites indicating the existence of possibly suitable data-sets, many of these sites are out-of-date or informally curated.
   - Formal experimental data-set registries are being actively implemented. These also need characterization and selection to find examples suitable for this project.
   - Stream based data-feeds of tagging events are particularly desirable, especially if the same sources have archives of their data-feeds.
3. Identify end-user communities that have an interest in curation of their subject areas. Although very disappointing, strong advice that the difficulty of progressing an ethnographic project through ethics committee process has caused me to design experiments using contemporary western user domains. I am still expecting to use members of some communities of interest - where groups have overlapping domains but divergent interests.
4. Perform preliminary processing of data-sets to refine technical approach and use that to inform further development of my thesis. I expect to iterate through this cycle several times.

5. Once algorithms for deriving categorizations latent in the test data-streams are performing adequately it will be feasible to begin on user-interface development and visualization. I expect to base the display and tagging interface for end-users on the The Virtual Museum of the Pacific, and iteratively evolve it from that base.
6. Once an end-user user-interface is working under informal testing, finalization of user-testing strategy should be developed. Resources such as Amazon Mechanical Turk will probably be used for scalable user-testing.
7. Finalize the system architecture for the experimental work-bench. At this stage of the project I should have sufficient knowledge of the characteristics of data-sets, user-iteration and analysis tools to perform this task with some confidence. Initially I am strongly motivated to use FOSS where practicable, and to implement the system so that it can be scaled by moving from a Linux-based workstation onto a cloud platform.
8. Complete the thesis and package the system for archiving.

7 Reflections

I don't regard my possible success as a sign of failure of others, rather I hope to answer a variant requirement with my own synthesis. Most of the work done with Ontology development has been very formal, which excludes most busy domain-experts from their development.

I hope to produce some interesting results which are a step on the way to generating useful, evolvable user-ontologies, and that the 'workbench' of tools I compose will be useful to others in the field, or a good starting point for further development.

8 Evaluation plan

There are three facets of evaluation for this project:

1. Subject domain usefulness - If a community of end-users tags objects from their domain of interest, they should recognize the extracted categories and rate them as useful
2. Ontologist usefulness - A working ontologist or curator should find the extracted categories useful in themselves, and in a standard format that they can use of in their normal working environment. They should also recognize features in the project's developed workbench that they would like to see as an improvement in their professional working environment.
3. Workbench deployment effectiveness - The running and deployment of the work-bench should be accessible to a researcher in ontology extraction who has a reasonable ability to maintain and configure a Linux workstation.
4. Demonstrate that relationships between annotators can be convincingly deduced by agent affiliation analysis from the tripartite graph of agent, object, and annotation. Agents would need some verifiable, relevant relationship whose description is independent of the tagging dataset.

5. Given a reasonably large tagging event stream from a particular community, determine how well computationally emergent ontologies from that dataset are received by that community.
6. Use temporal sliding windows to generate subsets of a large tagging event stream as input to observe if temporal category drift is observable.
7. Demonstrate computation that generates emergent ontologies incrementally in an environment where the whole data-set cannot be held in system RAM.

References

1. Bergman, M.: The sweet compendium of ontology building tools (Jan 2010), http://www.mkbergman.com/862/the-sweet-compendium-of-ontology-building-tools/
2. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. p. 113120 (2006), http://dl.acm.org/citation.cfm?id=1143859
3. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. In: Advances in Neural Information Processing Systems. p. 2003. MIT Press (2004)
4. Candy, L.: Practice based research: A guide. CCS Report: 2006-V1. 0 November p. 19 (2006)
5. Chandler, D.: Semiotics the Basics. Taylor & Francis, Hoboken (2007), http://public.eblib.com/EBLPublic/PublicView.do?ptiID=308502
6. Cimiano, P., Staab, S., Tane, J.: Automatic acquisition of taxonomies from text: FCA meets NLP. In: International Workshop & Tutorial on Adaptive Text Extraction and Mining held in conjunction with the 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of. p. 10 (2003)
7. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
8. Eklund, P., Goodall, P., Wray, T.: Information retrieval and social tagging for digital libraries using formal concept analysis. In: Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2010 IEEE RIVF International Conference on. p. 16 (2013)
9. Eklund, P., Goodall, P., Wray, T., Bunt, B., Lawson, A., Christidis, L., Daniels, V., Van Ollfen, M.: Designing the digital ecosystem of the virtual museum of the pacific. In: 3rd IEEE International Conference on Digital Ecosystems and Technologies. IEEE Press (2009)
10. Eklund, P., Goodall, P.J., Wray, T., Daniel, V., Van Olffen, M.: Folksonomy with practical taxonomy, a design for social metadata of the virtual museum of the pacific. In: Proceedings of the 6th International Conference on Information Technology and Applications. p. 112117 (2009)
11. Fischer, G.: Communities of interest: Learning through the interaction of multiple knowledge systems. In: Proceedings of the 24th IRIS Conference. p. 114 (2001)
12. Fu, W.T.: The microstructures of social tagging: a rational model. In: Proceedings of the 2008 ACM conference on Computer supported cooperative work. p. 229238 (2008), http://dl.acm.org/citation.cfm?id=1460600

13. Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubzy, M., Eriksson, H., Noy, N.F., Tu, S.W.: The evolution of protg: an environment for knowledge-based systems development. International Journal of Human-computer studies 58(1), 89123 (2003), http://www.sciencedirect.com/science/article/pii/S1071581902001271
14. Gruber, T.R., et al.: A translation approach to portable ontology specifications. Knowledge acquisition 5, 199199 (1993)
15. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web 5(1), 5–15 (Mar 2007), http://www.sciencedirect.com/science/article/B758F-4MYF67P-1/2/56984a3ddf4632bb98b722551cdb1151
16. Nth, W.: Handbook of semiotics. Indiana University Press (1995), http://books.google.com/books?hl=en&lr=&id=rHA4KQcPeNgC&oi=fnd&pg=PR9&dq=Handbook+of+Semiotics&ots=ddo1tWkS4f&sig=9NNalr_6O7bREpvCPhwQXB9lBn4
17. Oxford English Dictionary: "ontology, n.". (2004), http://www.oed.com/view/Entry/131551?redirectedFrom=Ontology
18. Rosch, E., Mervis, C., Gray, W., Johnson, D., Boyes-Braem, P.: Basic objects in natural categories. Cognitive psychology 8(3), 382439 (1976)
19. Rosch, E.: Principles of categorization. Concepts: core readings p. 189206 (1999), http://books.google.com/books?hl=en&lr=&id=sj1gczQ-7K8C&oi=fnd&pg=PA189&dq=%0D%0A1%0D%0APrinciples+of+Categorization%0D%0AEleanor+Rosch,+1978+&ots=NoqbGizR2u&sig=0WEAZVboo9yC0P8tnryxMC7DT3M
20. Sanborn, A.N., Griffiths, T.L., Navarro, D.J.: A more rational model of categorization. In: Proceedings of the 28th annual conference of the cognitive science society. p. 726731 (2006), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.3800&rep=rep1&type=pdf
21. Wenger, E.: Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press (Sep 1999)

NLP for Interlinking Multilingual LOD

Tatiana Lesnikova

INRIA & LIG, Grenoble, France
tatiana.lesnikova@inria.fr
http://exmo.inrialpes.fr/

Abstract. Nowadays, there are many natural languages on the Web, and we can expect that they will stay there even with the development of the Semantic Web. Though the RDF model enables structuring information in a unified way, the resources can be described using different natural languages. To find information about the same resource across different languages, we need to link identical resources together. In this paper we present an instance-based approach for resource interlinking. We also show how a problem of graph matching can be converted into a document matching for discovering cross-lingual mappings across RDF data sets.

Keywords: Multilingual Mappings, Cross-Lingual Link Discovery, Cross-Lingual RDF Data Set Linkage

1 Problem Statement

Due to the Resource Description Framework (RDF), the information on the Web can be turned from the unstructured mass into the structured data represented in the form of triples. The Linked Open Data (LOD) cloud containing billions of triples is constantly growing. Since data sets are created independently, there can be several Uniform Resource Identifiers (URIs) denoting the same entity across different RDF data sets. As a result, one needs to address the problem of entity resolution: identify and interlink the same entity across multiple data sources.

The RDF syntax is relatively simple and unambiguous: RDF = graph + identifiers (labels). This is what the identification of resources can be based on. However, this problem can become particularly difficult when there are multilingual elements in a graph as a simple string matching technique is doomed to fail. Hence, specific Natural Language Processing (NLP) techniques must be considered.

Our research problem is to find out methods for linking the same resource located in several RDF data sets and described in various natural languages and study the impact of available NLP techniques on the interlinking procedure.

2 Relevancy

Internet is a multilingual system, and we believe that it will continue to accommodate a diversity of natural languages despite the development of the Semantic Web. Even though there are many resources in English, some other languages occupy a decent portion of the Web space as well (see Internet world users by language statistics [1]). And we expect the necessity to tackle the multilinguality problem to persist. There are many resources that could be interlinked. At present, the number of languages [2] of RDF data sets amounts to 503.

The importance of cross-lingual mappings has been discussed in several works [1–3].

Recently a Best Practices for Multilingual Linked Open Data Community Group [3] has been created to elaborate a large spectrum of practices with regard to multilingual LOD.

Availability of the cross-lingual links is imperative for several neighboring research areas. For example, to overcome the problem of ontology heterogeneity, some research has been done on monolingual ontology integration based on instances interlinked by owl:sameAs [4]. If owl:sameAs links could be provided between instances expressed in different languages, other experiments on integrating underlying ontologies could be conducted.

The owl:sameAs links between instances can be also valuable in other applications such as Question Answering over multilingual structured knowledge-base [5] since a system can take advantage of the information presented in a language different from a language that is being queried.

Thus, the growing number of data sources in RDF format with multilingual labels and the importance of cross-lingual links for other Semantic Web applications motivate our interest in cross-lingual link discovery.

3 Related Work

The problem of searching for the same entity across multiple sources has been investigated in several research fields. In database community, it is known as instance identification, record linkage or record matching problem. In [6], the authors use the term "duplicate record detection" and provide a thorough survey on the matching techniques. Though the work done in record linkage is similar to our research, it does not contain cross-lingual aspect and RDF semantics.

In the field of Information Retrieval (IR), within the framework of the Cross-Language Evaluation Forum (CLEF)[4], the Web People Search Evaluation Campaigns (2007-2010)[5] focused on the Web People Search and person name ambiguity on Web pages and aimed at building a system which could estimate the

[1] http://www.internetworldstats.com/stats7.htm
[2] http://stats.lod2.eu/languages
[3] http://www.w3.org/community/bpmlod/
[4] http://www.clef-initiative.eu/
[5] http://nlp.uned.es/weps/weps-3

number of referents and cluster Web pages that refer to the same individual into one group. The research was performed on monolingual data.

Cross-lingual entity linking has been addressed in Knowledge Base Population track (KBP2011)[7] within the Text Analysis conference. The task is to link entity mentions in a text to a knowledge base (Wikipedia). If entity mentions are not in KB, they should be clustered into a separate group. Experiments were done both on monolingual (English) and cross-lingual data (Chinese to English). Authors in [8] used both language-independent and translation-based methods.

In contrast to the research outlined above, we aim at providing insights into the problem of cross-lingual interlinking from the point where data are already in RDF format, and we can vary different parameters in order to determine their impact on the interlinking operation.

In the Semantic Web, interlinking resources that represent the same real-world object and that are scattered across multiple Linked Data sets is a widely researched topic. Within the Data Interlinking track (IM@OAEI 2011), several interlinking systems have been proposed [9–13]. All of the systems were evaluated on monolingual data sets. Recent developments have been made also in multilingual ontology matching [14, 15].

To the best of our knowledge, there is no interlinking system specifically designed to link RDF data sets with multilingual labels.

4 Research Questions

The goal of our work is to provide methods to link interrelated resources across multilingual RDF data sets. For now, we restrict ourselves to owl:sameAs link [16] as it is a classical type of link that is usually established, and it is also important for tracking information about the same resource across different data sources. Given two RDF data sets with URIs and literals in different natural languages, the output will be a set of triples of type URI1 owl:sameAs URI2.

Our general *research question* is: To what extent is it possible to interlink data sets in different languages? To answer this question, within the framework that we describe in the Proposed Approach section, we need to explore which parameters influence this task. More specifically:

1. How to represent entities from RDF graphs?
   - What is the optimal distance for collecting language elements in traversal?
   - Is it necessary to preserve the structure of the graph in a virtual document by weighting the path length?
2. How to make entities described in different natural languages comparable?
   - What are the most appropriate Machine Translation techniques (rule-based, statistical, hybrid)?

- What is the impact of translating one language into another or pivot language?
- How does the output of similarity measures vary according to the context?

All these parameters will be studied with respect to specific contexts (language pairs, data set types, amount of textual data available). We also plan to experiment with graph matching techniques to see the difference with a translation-based approach. Apart from Machine Translation, we will explore techniques used for word alignment, thesaurus-based word sense disambiguation, multilingual document ranking, and mapping to multilingual lexicons.

5 Hypotheses

We introduce several hypotheses that we would like to test in our research.

1. If two URIs denote the same real-world object, the descriptions of the properties of this object should overlap with each other.
2. If descriptions are in different natural languages, then NLP techniques could help to decrease uncertainty across a set of resources.
3. If the descriptions of an entity overlap significantly, the similarity between them will be higher than between other entities.
4. If the degree of similarity depends on the available language context for each entity, then the more language data there are, the better will be the matching results.
5. If language data can be taken from two sources in RDF graph: property names and literals; then literals are more important since they are more informative.

6 Proposed Approach

Due to the presence of natural language terms in RDF graphs, we adopt a language-oriented approach.

The proposed approach includes several steps (see Figure 1).

1. Given two data sets with a resource representation in different natural language, extract language data for each URI. Thus, we create a "virtual" document for each URI.
2. Compare virtual documents in pairs from both sets.
3. Find the maximum similarity between two representations of the resource.
4. Establish an owl:sameAs link between the two most similar representations.

One should mind the following aspects of this approach:

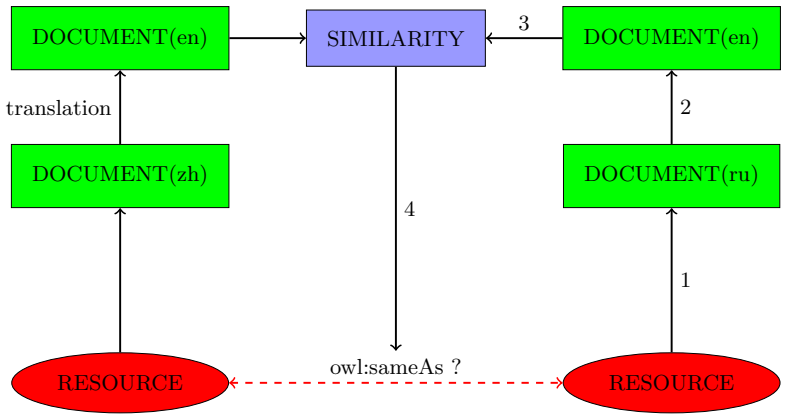


Fig. 1. Linking Process. Resources are described in Chinese and Russian languages and then translated into English.

– The idea of creating a "virtual" document has been employed in ontology matching [17]. The intuition of converting a graph into a document representation is that even though the taxonomy (structure) of graphs can be similar, the possibility to distinguish between two different things and identify the identical ones relies on their comparison. Thus, it is important to take into account lexical elements in a graph.

– Once we have documents representing resources, we need to decide how to define similarity between these resources. Similarity between documents can be taken for similarity between resources. Since we have documents in different languages, we can experiment with different types of Machine Translation (statistics-based, rule-based, hybrid). To estimate which strategy yields a better result, we will run our system by changing the translation component iteratively. Significant difference in results may signal which translation type is more beneficial. To enhance scalability, it would be interesting to translate the whole source corpus once and not to translate each label again and again. This would also allow for more contextual translation. The choice of translation techniques can also depend on the language combinations, for example, for rare languages, for which there does not exist enough parallel corpora, dictionary-based approaches might help.

– At the resource comparison step, it is important to reduce a number of possible comparisons for the sake of time-efficiency. For example, only comparisons between certain entity types are allowed. In case of using Supervised Machine Learning, the problem of training data is the most prominent one since there has been no official benchmark. And creating a generic training set for a heterogeneous amount of Linked Data seems very unrealistic. Then,

instead of training, it would be interesting to test clustering algorithms and find appropriate parameters for identity resolution.

- There are many techniques to compute similarity. A broad overview of them is given in [18]. We will use a vector space model [19] to represent terms in a "virtual" document as vectors of features. The choice of particular similarity measures is yet to investigate. When terms are in different languages, document similarity fails. Some similarity measures perform better on long texts. After transformation of "virtual" document into vectors, similarity metrics (e.g. Cosine, Euclidean) can be computed.

- A virtual document per URI shall contain language data in proximity to a given URI. The hypothesis is that the more textual data we have to characterize a resource, the easier it would be to identify the identical ones.

There are some complications as to textual data. Two scenarios are possible:

- URI can be looked up and the textual data extracted (as in case of DBpedia)
- No extra textual data are available per URI except the data in a graph itself.

To overcome this lack of context for a particular resource, we propose to browse a graph up to n+1 hops from the URI under investigation and collect data along the way. The data carriers are property names and literals. Thus, a virtual document for a particular URI will be the accumulation of data gathered during graph traversal.

This way of collecting a "profile" for a resource entails a question: does the difference of two graph structures affect the results of interlinking? On the one hand, taking into account the success of statistical machine translation based on statistical modelling and probabilities, the order of words is not always that important. It would be interesting to see whether it holds for RDF interlinking as well. On the other hand, we can try to preserve the order of collected properties and literals in a virtual document by putting weight for each language element. The further it is from the URI at question, the lower the weight. Term weight can be assigned by computing $termfrequency$ in a document or distribution of terms across a collection of documents known as $inverse\ documentfrequency$ (IDF). Terms that appear in few documents can be discriminative with regard to the rest of the documents. Combination of both TF x IDF is widely used in vector space models.

Once virtual documents are collected from both graphs, the documents will be compared and results evaluated.

7 Reflections

We believe that we can succeed in finding the solution for our research topic because we plan to put our research on a solid foundation and combine different methods to achieve the task. In traditional Web, there has been much

work done on multilingual NLP, i.e. language identification, machine translation, cross-language information retrieval. We are going to conduct series of experiments and see what works and how we can improve what does not work. This would allow us to preserve only the best practices and finally crystallize a solution to the problem. The author of this research proposal is also guided by the specialists in the domain that will contribute to the right choice of the research direction.

8 Evaluation

Evaluation means comparing the retrieved links against some reference. Standard measures usually serve for evaluation of an interlinking system (Precision, Recall, F-measure). The biggest challenge for evaluating our system is the absence of standard benchmark tests. As described in [20], there are several ways to go about this challenge. One of them would be to rely on the existing links between resources in DBpedia. This could be considered as a good alternative if not yet another hurdle: the existing interlanguage links can be inaccurate [21]. So, in our research we plan to experiment with different evaluation settings: we may experiment only with bi-directional links and/or study transitivity in order to ensure the correctness of test cases. The English, French, Russian versions of DBpedia [6] and Baidu Baike in Chinese [22] will be used for our experiments. We will also try to identify types of entities to focus on.

References

1. Gracia, J., Montiel-Ponsoda, E., Gômez-Pérez, A.: Cross-lingual Linking on the Multilingual Web of Data. In: Proc. of the 3rd Workshop on the Multilingual Semantic Web (MSW 2012) at ISWC 2012, Boston (USA), CEUR-WS ISSN 1613-0073, vol. 936 (2012)
2. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gômez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual Web of Data. Journal of Web Semantics, 11, 63–71 (2012)
3. Buitelaar, P., Choi, K.-S., Cimiano, P., Hovy, H. E.: The Multilingual Semantic Web (Dagstuhl Seminar 12362). Dagstuhl Reports 2(9), pp.15–94 (2012)
4. Zhao, L., Ichise, R.: Instance-Based Ontological Knowledge Acquisition. In: Proc.10th International Conference, ESWC 2013, Vol. 7882, pp.155–169. LNCS, Springer Berlin Heidelberg (2013)
5. Cabrio, E., Cojan, J., Gandon, F., and Hallili, A.: Querying multilingual DBpedia with QAKiS. In: Proc. 10th International Conference, ESWC 2013. Demo paper. Montpellier, France (2013)
6. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering. 19(1), 1–16 (2007)
7. Ji, H.,Grishman, R., Dang, H.T.: An Overview of the TAC2011 Knowledge BAse Population Track. In: Proc. Text Analytics Conference (TAC2011) (2011)

[6] http://dbpedia.org/About

8. Monahan, S., Lehmann, J., Nyberg, T., Plymale, J., Jung, A.: Cross-Lingual Cross-Document Coreference with Entity Linking. In: Proc. TAC2011 (2011)
9. Ngonga Ngomo, A.-C., Auer, S.: LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. IJCAI, pp. 2312–2317 (2011)
10. Nguyen, K., Ichise, R., Le, B.: SLINT: A Schema-Independent Linked Data Interlinking System. In: Proc. of the 7th International Workshop on Ontology Matching, pp.1–12 (2012)
11. Volz, J., Bizez, C., Gaedke, M., and Kobilarov, G.: Discovering and maintaining links on the web of data. In: Proc. of ISWC' 09, Springer-Verlag Berlin, Heidelberg, pp. 650–665, 2009.
12. Araújo, S., Hidders, J., Schwabe, D., Arjen, P. de Vries: SERIMI - Resource Description Similarity, RDF Instance Matching and Interlinking. CoRR, abs/1107.1104 (2011)
13. Niu, X., Rong, S., Zhang, Y., and Wang, H.: Zhishi.links results for OAEI 2011. In: Proc. of ISWC' 11 6th Workshop on Ontology Matching, pp. 220–227 (2011)
14. Meilicke, C., Trojahn, C., Sváb-Zamazal, O., Ritze, D.: Multilingual Ontology Matching Evaluation - a First Report on Using MultiFarm. In: Proc. of the 2d International Workshop on Evaluation of Semantic Technologies, pp.1–12, Heraklion, Greece (2012)
15. Meilicke, C., Castro, R.G., Freitas, F., van Hage, W.R., Montiel-Ponsoda, E., de Azevedo, R.R., Stuckenschmidt, H., Sváb-Zamazal O., Svátek, V., Tamilin, A., Trojahn, C., Wang, S.: MultiFarm: A Benchmark for Multilingual Ontology Matching. Journal of Web Semantics. 15, 62–68 (2012)
16. Halpin, H., Hayes, J. P.: When owl: sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In: Proc. of the Linked Data on the Web Workshop (LDOW2010), Raleigh, North Carolina, USA, April 27, 2010, CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-628/ldow2010_paper09.pdf
17. Qu, Y., Hu, W., Cheng, G.: Constructing virtual documents for ontology matching. In: Proc. of the 15th International Conference of World Wide Web, pp.23–31 (2006)
18. Euzenat, J. and Shvaiko, P.: Ontology Matching. Springer-Verlag, Heidelberg (2007)
19. Salton, G.: The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs, NJ (1971)
20. Lesnikova, T.: Interlinking Cross-Lingual RDF Data Sets. In: Proc. 10th International Conference, ESWC 2013, Vol. 7882, pp. 671-675. LNCS, Springer Berlin Heidelberg (2013)
21. Rinser, D., Lange, D., Naumann, F.: Cross-lingual entity matching and infobox alignment in Wikipedia. Information Systems, 38 (6), pp. 887-907 (2013)
22. Wang, Z., Wang, Z., Li J., Pan, J. Z.: Knowledge extraction from Chinese wiki encyclopedias. Journal of Zhejiang University - Science C 13(4): 268-280 (2012)

Optimizing RDF stores by coupling General-purpose Graphics Processing Units and Central Processing Units

Bassem Makni

Tetherless World Constellation, Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180
maknib@rpi.edu
http://tw.rpi.edu/

Abstract. From our experience in using RDF stores as a backend for social media streams, we pinpoint three shortcomings of current RDF stores in terms of aggregation speed, constraints checking and large-scale reasoning. Parallel algorithms are being proposed to scale reasoning on RDF graphs. However the current efforts focus on the closure computation using High Performance Computing (HPC) and require pre-materialization of the entailed triples before loading the generated graph into RDF stores, thus not suitable for continuously changing graphs. We propose a hybrid approach using General-purpose Graphics Processing Units (GPGPU) and Central Processing Units (CPU) in order to optimize three aspects of RDF stores: aggregation, constraints checking, and dynamic materialization.

1 Problem Statement

Social media graphs and streams fit naturally with the graph structure, and the Semantic Web offers the required platform to enrich, analyze and reason about social media graphs. However from our practical experience in storing, querying and reasoning about social media streams, we encountered the following problems:

1.1 Loading continuous stream of data

Performing constraints checking at load time, with *rdfs:domain* and *rdfs:range* for example, slows down adding new triples, and may result in a long queue of triples generated from the social media streaming API, waiting to be inserted into the graph. Turning off the constraints checking, at the other hand provides faster loading but may result in integrity violations in the graph.

1.2 SPARQL query performance

Running time of aggregation queries raises significantly with the size of the graph, however these aggregate functions are the most used to analyze the social media graph by queries like: "Return the number of persons tweeting about certain hashtag".
Moreover, SPARQL queries with large results, such as the ones extracting *mentions network* or returning millions of tweets with certain hashtags, for instance to calculate their sentiments, generate timeouts. Breaking down these large queries with offset and limit is not of big help as offset requires ordered triples and ordering big number of triples timeouts as well. This observation is partially supported by the Berlin SPARQL benchmark [1], and discussed by the Semantic Web community [2].

1.3 Large-scale reasoning

One of the major advantages of Semantic Web, is the ability to reason about the data. With the Linked Data movement gaining momentum, the amount of published RDF data is increasing significantly. And reasoning about large-scale RDF data, became a bottleneck. Recent research efforts tried to scale reasoning capability by parallelizing the closure computation. Thus the entailed RDF triples are materialized by forward-chaining reasoning. We call these approaches "offline parallelization", as the RDF graph needs to be generated offline on a cluster or super-computer before being loaded into an RDF store. However these offline approaches suffer from three drawbacks:

1. Forward-chaining generates a big number of entailed triples: BigOWLIM [3], for example, generated 8.43 billion implicit statements when loading 12.03 billion triples from the LUBM(90 000) [4] benchmark. However, not all the generated triples are usable at query time, which imply slowing down the RDF store by loading unusable implicit triples. That is one of the reasons why backward and hybrid chaining are more used in RDF stores. In Backward chaining, the entailments are computed at query time, and in hybrid chaining, a small number of rules is used in forward chaining and the rest are inferred at query time.
2. Offline parallelization is not practical for continuously changing graphs, as the closure needs to be recomputed on the supercomputer and loaded again in the RDF store whenever the graph changes. That may not be the case if the change is only by adding new triples, but deleting triples require necessarily the closure re-computation when the deleted triple affect the implicit generated triples.
3. They require HPC access and skills, as the user needs to compute the closure on a cluster or super-computer before being able to load and interact with the data via SPARQL.

We are particularly interested in reasoning about dynamic, i.e. continuously changing graphs, such as the social media graphs. So the need for an "online

parallelization" approach that is integrated into the RDF store and does not require previous processing of the graph.

2 Relevancy

Reasoning about large-scale and continuously changing RDF graphs, requires at the same time an online and parallel approach.
To the best of our knowledge, GPGPU are not exploited yet to optimize RDF stores reasoning and aggregation capabilities. Our goal is to design an RDF store that uses a hybrid GPGPU-CPU architecture, in order to support large-scale reasoning and optimize aggregation queries. The design should also provide parallelization capabilities in a transparent way, so the user will interact with the GPGPU-CPU RDF store in the same way he interacts with any CPU-only RDF store, without the need for HPC skills.

3 Related Work

The literature contains three classes of relevant works to our proposal: the first class is about closure computation parallelization, the second exploits GPGPU capabilities for graphs processing and the latter is about databases optimization using GPGPU.

3.1 Parallelization of closure computation

Three particular works will be of inspiration when designing the GPGPU-CPU RDF store, which are: Scalable distributed reasoning using MapReduce [5], Parallel materialization of the finite RDFS closure for hundreds of millions of triples [6] and The design and implementation of minimal RDFS backward reasoning in 4store [7]. The first two fit into the "offline parallelization" category, and the later uses parallel implementation of backward-chaining to support reasoning for the cluster based, 4store SPARQL engine. In the first paper the authors start from a naïve application of MapReduce on RDF graphs to adding a bunch of heuristics to optimize the parallelization of the closure computation. And in the second article, the authors describe an Message Passing Interface (MPI) implementation of their embarrassingly parallel algorithm to materialize the implicit triples.

3.2 Graphs processing on GPGPU

Unlike the early GPU units which were intended for graphical processing only, the General-purpose computing on graphics processing units can be used in more flexible ways via frameworks like Open Computing Language (OpenCL) [8] and Compute Unified Device Architecture (CUDA) [9]. The main difference between GPUs and CPUs is that GPUs launch a big number of light and slow threads

that run the same code in parallel, while CPUs run one fast process in serial. The parallel reduction, is a good illustration of the use of parallel threads on GPGPU to achieve high performance results. Figure 1 from the CUDA reduction white paper [10] illustrates the use of GPGPU to calculate the sum of an array.

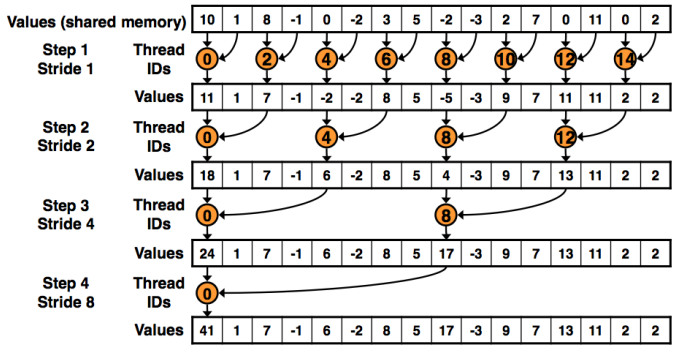


Fig. 1. Parallel Reduction: Interleaved Addressing

The literature contains many parallel algorithms that use GPGPU for fast sorting [11], DNA sequence alignment [12], etc. We focus on graphs processing algorithms, as they raise similar research problems to RDF processing, especially the data partition. In the early works on graph processing on the GPU, [13] achieved the performance of 1 second, 1.5 seconds and 2 minutes, respectively for a breadth-first search (BFS) single-source shortest path (SSSP), and all-pairs shortest path (APSP) algorithms on a 10 millions vertices graph. However these algorithms are limited to graphs with less than 12 millions vertices, as they load the whole graph into the GPGPU shared memory and they don't tackle the data partition issue. Recent research works break this scalability barrier by enabling graph data distribution. [14] use a multi-GPU architecture and balance the load of data between the different GPU devices.

3.3 Relational databases optimization on GPGPU

This research field tries to solve the similar problems that we are tackling for relational databases. In [15], the authors describe their GPU based, in-memory relational database GDB. [16] instead uses the traditional relational databases, and accelerates the SQL operations on a GPU with CUDA.

4 Research Questions

Unlike MPI and cluster based parallelization algorithms, which use the large memory available on every node, the GPGPU based algorithms are restrained by the relatively small memory of the GPU, and copying the data from the CPU to the GPU memory for parallel processing is an expensive operation. These constraints raise the following research questions when designing a GPGPU-CPU based RDF store:

Data partition One of the major constraints of GPU programming, is the small amount of dedicated graphical memory. Thus the need to swap data between the CPU memory or hard drives and the GPU memory. This need raises the research question of partitioning the RDF graph data, and the choice of the data slice that should be copied to GPU for parallel processing.

Dynamic computation As copying the data to GPU memory is time consuming, the overall speedup will depend on the parallel computation on the GPU and the data swap. If the speedup is not big, it is more efficient to compute on the CPU without the need to move the data. So the need to precompute this speedup in order to dynamically choose between CPU or GPU processing.

GPU caching In order to limit the number of memory fetches by the GPU, we need to adapt a caching mechanism to the small amount of memory, in order to maintain frequently used triples and rules in the shared memory.

Reasoning We need to select from the panoply of reasoning algorithms such as tableaux based, forward chaining, backward chaining etc. the one that is most adaptable to the GPU architecture, and can take benefit of the GPU parallelization.

5 Hypotheses

We design our GPGPU-CPU hybrid RDF store with the following hypothesis:

1. The graph structure is continuously changing, by adding new triples via social media streams. Triples can be deleted occasionally when the follower/friend relation is deleted for instance.

2. The user would interact with the RDF store in the same way he interacts with CPU based RDF stores, and the GPGPU parallelization should run in a transparent way.

6 Approach

We break down our approach into three steps related to the contributions we are promoting:

6.1 Optimizing SPARQL aggregation and ordering queries

In the first step, we are planning to integrate the CUDA reductions into the SPARQL runtime process. For example a query containing the *min* aggregator in Jena [17] is handled by the *AggMaxBase* class which go through the graph nodes and calculates sequentially the comparisons to the actual max. A GPGPU version will instead gather the whole sequence of bindings without any comparison and run a GPU reduction to get the max value in one shot.
In this step, we will get more familiar with the GPGPU programming and technical limitations, the first results of this step will allow us to tweak further steps plan.

6.2 Parallel constraints checking

One of the problems that we mentioned earlier in this paper, is the slow down of load speed, resulting from the constraints checking. We speculate that a GPGPU parallel version of constraints checking will optimize this phase and maintain the speed of inserting new coming triples from the social media streams. In this step we will get more familiar with lightweight reasoning on GPGPU.

6.3 Dynamic materialization on GPGPU

This step will be the core of the thesis work, as it raises the majority of the research questions discussed in this paper. We will use the lessons learnt from the previous steps in order to achieve parallel dynamic materialization of triples at the query time.

7 Reflections

GPGPU are being exploited by many research communities in order to provide cheap and fast parallelization of their algorithms. Successful results are achieved in bioinformatics, graphs processing, relational databases and other fields. We believe that exploiting GPGPU to optimize SPARQL queries and large RDF data processing is a fruitful research direction.
Though the problem of dynamic materialization on GPU, raises many research questions, we broke down our approach into intermediate warm-up steps before tackling this problematic.

8 Evaluation plan

We are planning to implement a GPGPU version of one, possibly two of the following RDF stores: Jena, Sesame [18] and Open Virtuoso [19]. In the evaluation phase, we plan to use SPARQL benchmarks such as the Berlin SPARQL benchmark [1] in order to compare the CPU only and the GPGPU-CPU hybrid version of each implemented RDF store.

As these benchmarks are not designed for streaming data, we will also use SRBench, the Streaming RDF/SPARQL Benchmark [20] which is specific for streaming RDF data. In [20], the authors propose a comprehensive set of queries in order to compare the main streaming engines available in the state of the art, namely $\text{SPARQL}_{\text{Stream}}$ [21], C-SPARQL [22] and CQELS [23].
Another possible benchmark to evaluate our approach, comes from the graphs processing community which is Graph500 [24].

Acknowledgements

I would like to express my deep gratitude to my advisor Prof. James Hendler and to Prof. Deborah McGuinness for their guidance in this research proposal. This work was supported in part by the DARPA SMISC program.

References

1. Bizer, C., Schultz, A.: The berlin sparql benchmark. International Journal on Semantic Web and Information Systems (IJSWIS) **5**(2) (2009) 1–24
2. semanticweb.com: Set of real-world sparql benchmark queries. `http://answers.semanticweb.com/questions/1847/set-of-real-world-sparql-benchmark-queries` (2010)
3. Kiryakov, A., Bishop, B., Ognyanoff, D., Peikov, I., Tashev, Z., Velkov, R.: The features of bigowlim that enabled the bbcs world cup website. In: Workshop on Semantic Data Management SemData VLDB. (2010)
4. Guo, Y., Pan, Z., Heflin, J.: Lubm: A benchmark for owl knowledge base systems. Web Semantics: Science, Services and Agents on the World Wide Web **3**(2-3) (2011)
5. Urbani, J., Kotoulas, S., Oren, E., Van Harmelen, F.: Scalable distributed reasoning using mapreduce. In: The Semantic Web-ISWC 2009. Springer (2009) 634–649
6. Weaver, J., Hendler, J.A.: Parallel materialization of the finite rdfs closure for hundreds of millions of triples. In: The Semantic Web-ISWC 2009. Springer (2009) 682–697
7. Salvadores, M., Correndo, G., Harris, S., Gibbins, N., Shadbolt, N.: The design and implementation of minimal rdfs backward reasoning in 4store. In: The Semanic Web: Research and Applications. Springer (2011) 139–153
8. Munshi, A., et al.: The opencl specification. Khronos OpenCL Working Group **1** (2009) l1–15
9. Nvidia, C.: Programming guide (2008)
10. Harris, M.: Optimizing parallel reduction in cuda. NVIDIA Developer Technology **6** (2007)
11. Merrill, D.G., Grimshaw, A.S.: Revisiting sorting for gpgpu stream architectures. In: Proceedings of the 19th international conference on Parallel architectures and compilation techniques, ACM (2010) 545–546
12. Trapnell, C., Schatz, M.C.: Optimizing data intensive gpgpu computations for dna sequence alignment. Parallel Computing **35**(8) (2009) 429–440
13. Harish, P., Narayanan, P.: Accelerating large graph algorithms on the gpu using cuda. In: High Performance Computing–HiPC 2007. Springer (2007) 197–208
14. Merrill, D., Garland, M., Grimshaw, A.: Scalable gpu graph traversal. In: Proceedings of the 17th ACM SIGPLAN symposium on Principles and Practice of Parallel Programming, ACM (2012) 117–128

15. He, B., Lu, M., Yang, K., Fang, R., Govindaraju, N.K., Luo, Q., Sander, P.V.: Relational query coprocessing on graphics processors. ACM Transactions on Database Systems (TODS) **34**(4) (2009) 21
16. Bakkum, P., Skadron, K.: Accelerating sql database operations on a gpu with cuda. In: Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units, ACM (2010) 94–103
17. McBride, B.: Jena: Implementing the rdf model and syntax specification. (2001)
18. Broekstra, J., Kampman, A., Van Harmelen, F.: Sesame: A generic architecture for storing and querying rdf and rdf schema. In: The Semantic WebISWC 2002. Springer (2002) 54–68
19. Erling, O., Mikhailov, I.: Rdf support in the virtuoso dbms. In: Conference on Social Semantic Web. Volume 113. (2007) 59–68
20. Zhang, Y., Duc, P.M., Corcho, O., Calbimonte, J.P.: Srbench: a streaming rdf/sparql benchmark. In: The Semantic Web–ISWC 2012. Springer (2012) 641–657
21. Calbimonte, J.P., Corcho, O., Gray, A.J.: Enabling ontology-based access to streaming data sources. In: The Semantic Web–ISWC 2010. Springer (2010) 96–111
22. Barbieri, D.F., Braga, D., Ceri, S., Valle, E.D., Grossniklaus, M.: Querying rdf streams with c-sparql. ACM SIGMOD Record **39**(1) (2010) 20–26
23. Le-Phuoc, D., Dao-Tran, M., Parreira, J.X., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. In: The Semantic Web–ISWC 2011. Springer (2011) 370–388
24. Murphy, R.C., Wheeler, K.B., Barrett, B.W., Ang, J.A.: Introducing the graph 500. Cray Users Group (CUG) (2010)

Television meets the Web: a Multimedia Hypervideo Experience

José Luis Redondo García, Raphaël Troncy

EURECOM, Sophia Antipolis, France,
{redondo, raphael.troncy}@eurecom.fr

Abstract. Nowadays, more and more information on the Web is becoming available in a structured and easily searchable way. Multimedia and particularly television content has to adopt those same Web principles in order to make the consumption of media items and Web browsing seamlessly interconnected, no matter if the content is coming live from a local broadcaster or from an online video streaming service. This research proposes the creation of a methodology for representing, searching and interlinking multimedia items with other existing resources. This model has to be easily extensible and Web compliant in order to make the generated annotations available to the outside world and reused in other similar applications.

Keywords: Hypervideo, Television, Multimedia Ontology, NER, Second Screen

1 Problem Statement

The amount of multimedia content available is huge: there are billions of pictures and videos spread in very different ecosystems that keep growing. Since each of those ecosystems has is own peculiarities, it is not easy to identify which media items are relevant in a particular context, and how they can be effectively consumed or re-used by users. Television content constitutes a subset of this multimedia world where the presence of isolated ecosystems becomes even more obvious. Most of the television providers are media silos ruled by their own idiosyncrasy: broadcasters such as the German RBB or the French TF1 groups are the owners of tons of television content that are not conveniently exposed and interlinked with the rest of the world.

Other open issue is the way those media items are described. In most of the cases, the content is considered like an unitary piece that does not need to be further fragmented. However, in many situations, a more fine-grained decomposition of the media resource is needed, in order to point to particular fragments where something of interest is happening. The broad variety of non-interoperable standards for representing those descriptions, such as TV-Anytime[1] or MPEG-7[2], has been largely recognized.

[1] http://tech.ebu.ch/tvanytime
[2] http://mpeg.chiariglione.org/standards/mpeg-7

Finally there is a need of complementing seed video programs with additional resources on the Web, such as domain related web sites, encyclopedic sources or even conversation happening on social platforms. Users' status updates and tweets enable people to share their activities, feelings and emotions opening a window to a world of fresh information that can successfully illustrates what is being broadcasted in the television screen.

2 Relevancy

Multimedia content is rapidly increasing in scale and ubiquity but it still remains largely poorly indexed and unconnected with other related media from other sources. The current state of the TV domain clearly reflects this fact: there are no clear approaches for going further than a simple PC-like browsing experience on a full screen. Users wish to have new functionalities such as getting access to information not explicitly present in the television content itself, like browsing from a local news show to an open government data portal about a particular location in order to understand voting patterns, or learning more about animals and plants shown in a nature documentary without leaving that show.

The information available in social platforms is growing and becoming more and more attached to television programs. Filtering this massive amount of data and trying to make sense out of it is an extremely challenging task due to the heterogeneity and dynamics of the information. Such an analysis is, however, very valuable for explaining or illustrating what is going on in a video using similar audiovisual content or encyclopedia knowledge, for completing missing information, or for providing other users' point of view about the same fact or topic.

3 Related Work

Several approaches for describing multimedia content in general and television content in particular have been proposed. One of the most relevant ones is the model proposed by the British broadcaster, the BBC Programmes Ontology[3]. This ontology includes concepts about various aspects of a television program such as series, brands, episodes, etc. In [1], the authors describe how the BBC is working to integrate data and linking documents across their domains by using Semantic Web technology. The latest version of schema.org includes also classes related to the TV domain, such as `TVSeries`, `TVSeason` and `TVEpisode` which belong to the SchemaDotOrgTV[4] proposal. This set of classes include some improvements over the BBC model, like a better relationship between episodes, series and seasons or the addition of clips and broadcast services.

Those attempts consistently consider the television program as a whole and do not consider its sub-parts or fragments. The possibility of working with pieces

[3] `http://purl.org/ontology/po/`
[4] `http://www.w3.org/wiki/SchemaDotOrgTV`

of different granularities is a crucial requirement for implementing true hypervideo systems. There are standards that use non-URI based mechanisms to identify parts of a media resource, such as MPEG-7 or the Synchronized Multimedia Integration Language (SMIL)[5]. In the group of URI-based approaches, temporalURI[6] was the first to define media fragments using the query parameter in a URI. The W3C Media Fragment Working Group has edited the Media Fragment URI 1.0 specification[7], which defines a hash URI syntax for identifying and referencing fragments of audiovisual content in the Web. Some systems, such as Synote [3], rely on these media fragments for representing information about subtitles and entities in videos. In [2], we have used media fragments and entities for classifying videos from Dailymotion and YouTube.

Crawling social platforms for enriching a media resource has been proposed in several work. Liu *et al.* combine semantic inferencing and visual analysis to automatically find media to illustrate events [4]. Visual, temporal and spatial similarity measures are used for attaching photo streams with events in [9]. We developed a generic media collector for retrieving media items shared on social platforms [6]. Extra insights about the facts and events illustrated in the media items collected are inferred by performing non-supervised clustering and labeling processes on the result set.

4 Research Questions

The first challenge is to find an appropriate ontology model for correctly representing the metadata about a particular media resource. We hypothesis that Media Fragment URI can be used to identify part of media content, but it is still a key topic of discussion how the relationship between media fragments and a media resource should be represented and what are the implications for the annotations.

The second challenge is how to enrich and improve the available metadata by retrieving extra information from the Web and media items published in social platforms. Some questions are: (i) what are the anchors or properties inside a media resource that can best describe a particular fragment, (ii) what is the best way to materialize links between different media fragments, and (iii) how to formalize the context surrounding an annotation including its provenance, the motivation that leads to the creation of the annotation, etc., in order to better support search and hyperlink operations.

Finally, we aim to investigate how this enriched television content with data from the web interconnected to its subparts should be consumed and displayed to the end-user. The presence of multiple links to other resources opens a window to a great variety of new television applications where a much minimalist interaction and interface design should be implemented.

[5] `http://www.w3.org/AudioVideo/`
[6] `http://annodex.net/TR/draft-pfeiffer-temporal-fragments-03.txt`
[7] `http://www.w3.org/TR/media-frags/`

5 Hypotheses

This research proposed the use of a semantic graph metadata representation for implementing innovative hypervideo systems. The resulting RDF data is flexible enough to include different types of content description in a structured way: it can be completed with information from external resources, it naturally supports links with other pieces of content, and its web nature enables to bring hypermedia experience to the TV field.

A video program can be decomposed into segments, either automatically or manually, in order to create a hierarchy (structure) of media fragments which can be further indexed and semantically described with resources from the Web of Data. In our hypothesis, those anchors are Named Entities [8] spotted by different extractors used on timed texts coming with a media resource. Those entities represent a bridge between the audiovisual content and related information in the Web of Data that is potentially relevant for the viewer. By filtering and ranking those entities, the important parts of the video can be identified, modifying and further adjusting an existing segmentation result obtained via a visual analysis. Finally, the set of relevant entities inside a particular Media Fragment becomes an appropriate way of characterizing it, which allows to infer how similar a part of the video is to other fragments from other multimedia resources. By making explicit relationships between analogous media fragments, the expected hypermedia experience can be effectively created.

6 Approach

This sections shows a first implementation of the proposed hypotheses made in the context of the LinkedTV project[8], which will be further refined during the subsequent phases of this doctoral research.

RDF conversion. In a first step, some legacy metadata or some results coming from automatic multimedia analysis processes (e.g. face detection, shot segmentation, scene segmentation, concept detection, speaker identification, automatic speech recognition, etc.) are converted into RDF and represented according to the LinkedTV Ontology[9]. We have developer a REST API service named *tv2rdf*[10] to perform this operation. The video content is structured into parts, with different degrees of granularity, identified using Media Fragments URIs. For better classifying those different levels of segmentation, the LinkedTV ontology includes classes such as `Chapter`, `Scene` or `Shot`.

Those instances of the *ma:MediaFragment* class are anchors where entities will be attached in the following serialization step. The media fragment generation introduces a very important level of abstraction that opens many possibilities when annotating certain parts of the analyzed videos and makes possible to associate to fragments other metadata with temporal references. The underlying

[8] `http://www.linkedtv.eu/`
[9] `http://data.linkedtv.eu/ontologies/core`
[10] `http://linkedtv.eurecom.fr/tv2rdf`

model also relies on well-known ontologies such as the *The Open Annotation Core Data Model*[11], the *Ontology for Media Resources*[12] and the *NERD ontology*. A schema of the ontology is depicted in Figure 1.

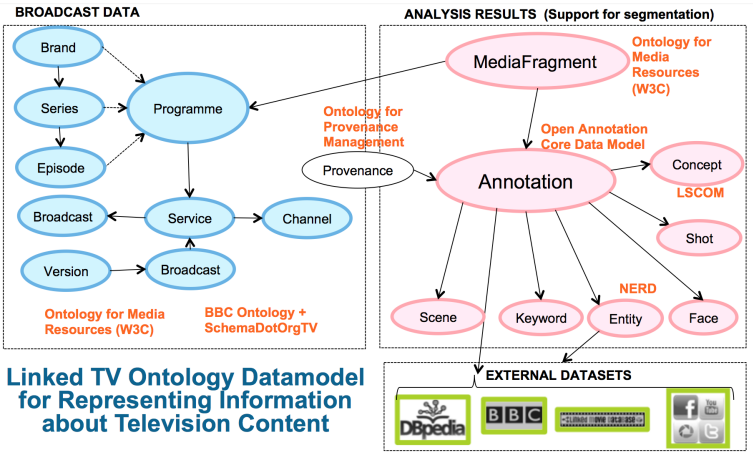


Fig. 1. LinkedTV Ontology.

The listing below corresponds to the description (Turtle serialization) of a media fragment from the *Tussen Kunst en Kitsch* show. The temporal references are encoded using the NinSuna Ontology[13]. The fact that one media fragment belongs to a larger media resource is made via the property `ma:isFragmentOf`.

```
<http://data.linkedtv.eu/media/e2899e7f-67c1-4a08-9146-5a205f6de457#t
    =1492.64,1504.88>
      a  nsa:TemporalFragment , ma:MediaFragment ;
      nsa:temporalStart "1492.64"^^xsd:float ;
      nsa:temporalEnd "1504.88"^^xsd:float ;
      nsa:temporalUnit "npt" ;
      ma:isFragmentOf <http://data.linkedtv.eu/media/e2899e7f-67c1-4a08
          -9146-5a205f6de457> .
```

Broadcasters generally provide basic legacy metadata related to the TV content such as EPG information (title, description, tags, channel, category, duration, language) and subtitles. Those items are also included in the RDF graph during the serialization process.

Name Entity Extraction. Once the RDF graph is built, some nodes are further interlinked with the Linked Open Data Cloud. Named entity extractors are used over the transcripts of the TV content (either the subtitles of the television program or the automatic speech recognition (ASR) results). The tv2rdf REST service launches this task by relying on a *NERD Client*, part of the

[11] `http://www.openannotation.org/spec/core`
[12] `http://www.w3.org/ns/ma-ont`
[13] `http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna`

NERD[14] framework. A multilingual entity extraction is performed over the video transcript and the output result is a collection of entities that are temporally related to a video. The entities are classified using the core NERD Ontology[15] and attached to the right Media Fragment according to their temporal appearance.

Both Dublin Core[16] and LinkedTV properties are used in order to specify the entity label, confidence and relevance scores, the name of the extractor used, the entity type and the disambiguation URI for this entity (for example, a resource from DBPedia). This entity extraction process can be extended to be also applied over other textual resources such as users comments or notes from the publisher.

Enrichment. In a third step, the named entities extracted in the previous step are used to trigger the enrichment process that consists in retrieving additional multimedia content that can illustrate what is shown or discussed in a seed television program. The logic for accessing the external datasets where this information can be collected is implemented inside the LinkedTV REST service MediaCollector[17]. MediaCollector gets as input the label of entities spotted by NERD and provides as result a list of media resources (photos and videos) grouped by source [7]. Those sources include mainly social platforms such as Twitter or YouTube but also domain-related web sites provided as white list of content that should be mined. When serializing the information into RDF, every item returned by the MediaCollector is represented as a new `ma:MediaResource` instance according to the Ontology for Media Resources. The entity used as input in the media discovery process is linked to the retrieved items through an `oa:Annotation` instance from the Open Annotation Ontology.

Search and Hyperlinking. Entities and related media items can then be used to describe and illustrate a particular media fragment of the television program. Given some input parameters from the viewer and analyzing the entities that are relevant for a video fragment, it is possible to filter out the ones that can be potentially interesting for the user. When different media fragments share similar named entities, they can be explicitly interrelated through the creation of hyperlinks that allow the user to navigate from one multimedia content to the other.

7 Reflections

We aim to publish the RDF metadata following the linked data principles. Hence, the resulting RDF graph can not only be stored and queried to enable data visualization such as a second screen application [5], but it can also be re-used by other RDF consuming applications. Once the metadata about a particular content has been gathered, serialized into RDF, and interlinked with other resources in the Web, it is better suited to be used in the subsequent consumption phases such as an editorial review or a data visualization. The creation of a hierarchy

[14] `http://nerd.eurecom.fr/`
[15] `http://nerd.eurecom.fr/ontology/nerd-v0.5.n3`
[16] `http://dublincore.org/documents/2012/06/14/dces`
[17] `http://linkedtv.eurecom.fr/api/mediacollector/`

of media fragments with different levels of granularity provides a flexible model for allowing different interpretations of the data depending on the user or the particular context.

Different entities spotted within the same media fragment of a video resource can be considered simultaneously for obtaining new insights about what is happening in that part of the media resource. For example, let's imagine that an entity corresponding to a particular painting has been spotted inside a media fragment, while another entity known to be an artist in DBpedia is also spotted nearby. It would then be possible to infer that this person is the author of the painting with some confidence level, or at least, that this person is somehow related with it. Similar deductions can be done by relying on other annotations in the model such as keywords and LSCOM concepts (i.e. visual concepts as popularized by the TRECVID benchmarking campaign).

8 Evaluation plan

The cooperation with the AVRO (via Sound and Vision) and RBB broadcasters as partners of the LinkedTV project opens many possibilities. Apart of having more material to be processed, their viewers and editors can potentially test new features and provide feedback about the quality of the annotations and the usefulness of the enrichment process described in this paper.

The evaluation of the entire approach will be based on three complementary dimensions. First, the accuracy of the named entity extraction process. Are entities correctly spotted and disambiguated? This will be evaluated by applying standard metrics from the NLP domain. Second, the adequacy of the media fragment temporal boundaries and their relevance for the spotted entities. What is the right temporal window to consider around particular entities? Are they meaningful according to the story being told in the video? We aim to compare those results with with ground truth annotations manually generated by users. Finally, we will measure the precision and recall of the search and hyperlinking operations over the generated Media Fragments: for a particular search term given by a user, are the media fragments retrieved relevant? Are the links between media fragments interesting from a user point of view? The evaluation of this part is probably the most subjective and complex one but we plan to rely on standard datasets and in particular on the MediaEval Search and Hyperlinking Task[18] that has exactly this goal.

We have already performed some very preliminary evaluations. We computed some basic statistics about the number of named entities per NERD type and the number of media fragments in a 55 minutes episode of the show *Tussen Kunst en Kitsch* from the Dutch broadcaster Avro (Tables 1 and 2).

[18] `http://www.multimediaeval.org/mediaeval2013/`

Table 1. Number of entities per type

NERD type	Entities
Person	37
Location	46
Product	3
Organization	30

Table 2. Number of MediaFragment's

Serialized Item	MediaFragment
Shots&Concepts	448
Subtitles	801
Bounding Boxes	4260
Spatial Objects	5

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the project LinkedTV (GA 287911).

References

1. G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media Meets Semantic Web — How the BBC Uses DBpedia and Linked Data to Make Connections. In *6th European Semantic Web Conference (ESWC'09)*, pages 723–737, Heraklion, Crete, Greece, 2009.
2. Y. Li, G. Rizzo, J. L. R. García, R. Troncy, M. Wald, and G. Wills. Enriching media fragments with named entities for video classification. In *1st Worldwide Web Workshop on Linked Media (LiME'13)*, pages 469–476, Rio de Janeiro, Brazil, 2013.
3. Y. Li, M. Wald, T. Omitola, N. Shadbolt, and G. Wills. Synote: Weaving Media Fragments and Linked Data. In *5th Workshop on Linked Data on the Web (LDOW'12)*, Lyon, France, 2012.
4. X. Liu, R. Troncy, and B. Huet. Finding Media Illustrating Events. In *1st ACM International Conference on Multimedia Retrieval (ICMR'11)*, Trento, Italy, 2011.
5. V. Milicic, J. L. R. García, G. Rizzo, and R. Troncy. Grab your Favorite Video Fragment: Interact with a Kinect and Discover Enriched Hypervideo. In *11nd European Interactive TV Conference (EuroITV'13), Demo Track*, Como, Italy, 2013.
6. V. Milicic, G. Rizzo, J. L. R. García, R. Troncy, and T. Steiner. Live Topic Generation from Event Streams. In *22nd International World Wide Web Conference (WWW'13), Demo Track*, pages 285–288, Rio de Janeiro, Brazil, 2013.
7. G. Rizzo, T. Steiner, R. Troncy, R. Verborgh, J. L. R. García, and R. V. de Walle. What Fresh Media Are You Looking For? Retrieving Media Items from Multiple Social Networks. In *1st International Workshop on Socially-Aware Multimedia (SAM'12)*, Nara, Japan, 2012.
8. G. Rizzo and R. Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France, 2012.
9. J. Yang, J. Luo, J. Yu, and T. S. Huang. Photo stream alignment for collaborative photo collection and sharing in social media. In *3rd ACM International Workshop on Social Media (WSM'11)*, pages 41–46, Scottsdale, Arizona, USA, 2011.

Explaining data patterns using background knowledge from Linked Data

Ilaria Tiddi

Knowledge Media Institute, The Open University, United Kingdom
ilaria.tiddi@open.ac.uk

Abstract. When using data mining to find regularities in data, the obtained results (or patterns) need to be interpreted. The explanation of such patterns is achieved using the background knowledge which might be scattered among different sources. This intensive process is usually committed to the experts in the domain. With the rise of Linked Data and the increasing number of connected datasets, we assume that the access to this knowledge can be easier, faster and more automated. This PhD research aims to demonstrate whether Linked Data can be used to provide the background knowledge for pattern interpretation and how.

Keywords: Linked Data, Data Mining, Knowledge Discovery, Data Interpretation

1 Problem Statement

Knowledge Discovery in Databases (KDD) can be defined as the process of detecting hidden patterns and regularities in large amounts of data [**?**]. To be interpreted and understood, these patterns require the use of some background knowledge, which is not always straightforward to find. In most real world contexts, providing the background knowledge is committed to the experts, whose work is to analyse the results of a data mining process, give them a meaning and refine them. The interpretation turns out to be an intensive and time-consuming process, where part of knowledge can remain unrevealed or unexplained.

Our problem is illustrated with a real-world example we will use throughout this paper. The *Reading Experience Database* (RED)[1] is a record of people's reading experiences, including metadata regarding the reader, author, and book involved in the experience as well as its date and location. Several kinds of data mining processes could be applied on such a dataset. Here, for example, we look at how people can be clustered based on the similarity of what they read. Considering one such cluster, the question then becomes: "is there a reason why these people read the same kind of books?" and "where and how to find this information?". Given for instance a cluster of people having extensively read Jane Austen, an expert might consider it pertinent to point out that many of them are Anglican women, since, for a number of reasons, Jane Austen was more significantly popular with this particular audience.

Our hypothesis is that the access to the required background knowledge (in our example, that readers are Anglican and female) can be made easier with

[1] http://www.open.ac.uk/Arts/RED/index.html

Linked Data[2]. In fact, while in the documentary Web the information used to be hard to detect, hidden or even unreachable, the rise of Linked Data has made possible to directly access it. In the last decades, people have been putting efforts together in order to openly publish and link their knowledge in the form of domain-specific concepts and relationships. While Tim Berner's Lee's "Web of Data" [**?**] is still evolving and taking form, this structure and interoperability of data can already be exploited for the knowledge interpretation process.

2 Relevancy

In many real-world domains, background knowledge plays a central role for the analysis of trends or common behaviours. Generally, this knowledge is provided by experts interpreting the results and assisting the Knowledge Discovery process, which proves to be intensive and time-consuming.

In **Business Intelligence** (BI), the regularities emerging from raw data using data analytics are explained and transformed into meaningful information by an expert for business purposes, such as decision making or predictive analytics.

The young field of **Learning Analytics** aims at identifying trends and patterns from educational data using data mining, BI and Human-Computer Interaction techniques. The explanation of behaviours is crucial to assist people's learning, help teachers to support students, improve courses, as well as support the staff in planning and taking decisions.

In **Medical Informatics**, computer technologies are applied to process medical information. The explanation of trends and anomalies might come from some external knowledge, which the expert might not be aware of. A typical example is the environmental changes affecting the spread of diseases.

The analysis of data is also central in the field of **Humanities**, where researchers attempt to explain facts by finding hidden connections with some external sources. The RED example of the paper comes from this field.

These examples show on the one hand how background knowledge is required to explain the regularities in data, and on the other hand how this explanation can sometimes come from very different domains, not related to each other.

3 Related Work

While ontologies have been widely explored in the data mining context since the early 2000s, the last years have seen an increasing number of researches aiming at exploiting the potential of Linked Data. The overall idea behind the two trends is to exploit the datasets' structure and semantics and combine them with the Machine Learning algorithms to produce more accurate results. Earlier works proposed the use of ontologies as a support for data preparation [**?**,**?**,**?**] or to constrain the algorithms search [**?**,**?**,**?**,**?**]. Linked Data-driven approaches can be found in [**?**,**?**,**?**]. On the other hand, few works [**?**,**?**] had been addressed so far on using ontologies to assist the interpretation of the results. Recently, the idea has been considered in [**?**,**?**], where the authors stress the importance of

[2] http://linkeddata.org/

capturing useful knowledge from ontologies to reduce the user's workload for the interpretation process. While the motivation of both these works and ours is to combine ontologies and data mining in view of a complete virtuous KDD process, we also intend to further the use of Linked Data. This idea can be found in [**?**], where Linked Data are used to understand the results of a Sequence Pattern Mining process in the context of Learning Analytics. Linked Data are here only a navigation support to the user (who can easily visualise the results), while the interpretation is still based on his previously acquired expertise.

Ontologies for hypothesis generation have been treated in the clinical domain (see survey in [**?**]), and combined with Logic Programming in the fields of Description Logic Programming [**?**,**?**], as well as in the Onto-Relational Learning domain [**?**]. Particularly, this last approach exploits the unary and binary predicates of ontologies, to provide a strong background knowledge and combine it with Inductive Logic Programming in order to produce rules or hypotheses from observations.

4 Research Questions

The main research question we address is this work is: "*how do we explain patterns in data using the background knowledge from Linked Data?*". If, on one side, this "explanation" means the generation of some *hypotheses* (or rules) interpreting the data patterns, on the other side, these hypotheses should rely on some background knowledge that needs to be somehow retrieved, and we assume that at least some of it might be available through Linked Data. To answer this, we articulated our space in a specific set of subquestions, possible solutions and expected risks, which are illustrated below.

Q1 – Finding the data. Our first question is *how to find* the right background knowledge in Linked Data. This is our major question, and is articulated in:

1. *Dataset selection.* Does the Linked Data cloud contains the right datasets describing our data? Where and how to find them?
2. *Data detection.* Once we have found the datasets, how to detect the correct data into them? Do the data have enough information? In other words, how do we find the *correct pieces of knowledge*, in terms of predicates about our data?

Initial Solution. The question here concerns the exploration of the Linked Data cloud and the knowledge herein represented. While technical solutions such as the CKAN API[3], the Semantic Web indexers[4] or the SPARQL endpoint lists[5] are already popular in the community, our objective is to automatise the process of selecting the important bits of information required for the explanation. Whether we choose a top-down approach, where the search space is first defined by deeply analysing the datasets and then narrowed using the initial data to detect the salient bits of information for hypotheses generation, or a bottom-up

[3] http://docs.ckan.org/en/latest/api.html
[4] such as Sindice: http://sindice.com/
[5] http://www.w3.org/wiki/SparqlEndpoints

approach, that exploits the initial data to iteratively add pieces of (Linked Data) background knowledge to produce more and more refined hypotheses, the key of the process are the available connections in the Linked Data cloud. Exploiting such connections to make emerge underlying knowledge in order to maximise the automatisation of this selection process will be our major contribution.
Expected risks. The search for background knowledge could be unsuccessful as the patterns might not be described enough in Linked Data (*lack of information* or *lack of datasets*).

Q2 – Generating the hypotheses. Assuming that the background knowledge about the data has been found, we will have to answer the question: *how do we use it to explain* the data patterns. What kind of mechanisms can generate explanations, that we previously called hypotheses?
Initial Solution. We identified as a possible solution the use of Inductive Logic Programming to produce hypotheses from both data patterns and Linked Data background knowledge.
Expected risks. The chosen mechanism to generate explanations might not be scalable and might lead to computational problems (*data deluge*).

Q3 – Evaluating the hypotheses. Once the hypotheses have been generated, the last questions is: how do we know that they are *good*? That is, what is the significance of a rule? This evaluation step is also two-folded:

1. *Hypotheses evaluation.* Which are the criteria to assess the interestingness of a hypothesis?
2. *Method evaluation.* How do we evaluate that our method is efficient when compared to those of the domain(s) experts?

Finally, can the evaluation method affect the data selection? Can a hypothesis help in pruning the selected data, and support the Knowledge Discovery process?
Initial Solution. Currently, we are exploiting the ILP evaluation measures to score the significance of a hypothesis. However, we are aware that this preliminary solution will need to be further investigated. We also intend to investigate genetic algorithms to verify if the evaluation method can affect the data selection.
Expected risks. A clear evidence for some of the generated rules could be missing (*lack of background knowledge*). Moreover, some of the hypotheses might iteratively require a new piece of knowledge to explain the patterns (*recursion issue*).

5 Hypothesis

Our hypothesis is: "*Linked Data can be used as background knowledge to explain data patterns*". The main idea is that using Linked Data as background knowledge will reduce the efforts put into explaining the data patterns. Assuming this, Knowledge Discovery can leverage Linked Data as they will assist the experts and reduce their commitment into the KDD process, as explained below.

Time gaining. The expert will require **less time** to explain patterns. The connections between datasets of different areas will make emerge new information for the explanations requiring external knowledge.
Efficiency. Linked Data will show the expert the information which is not from his domain. Our method can be **more efficient** than a group of experts.
Completeness. The expert can be less specialised as Linked Data can bring the missing information, in order to have a **more complete** explanation.

6 Approach

The approach is structured according to our research questions (see also Fig. **??**).
1. *Data Selection.* Assuming some patterns obtained from a data mining process (clusters, association rules, sequence patterns...), we search in the Linked Data cloud information about the data in the patterns.
2. *Hypotheses Generation.* We use Inductive Logic Programming to represent both the data patterns and the Linked Data information, and generate hypothesis from them.
3. *Hypotheses Evaluation.* We evaluate the hypotheses in order to rank them and select the best rules. These are presented to the experts for interpretation, but also used to refine the data selection of the first step and to start a new cycle.

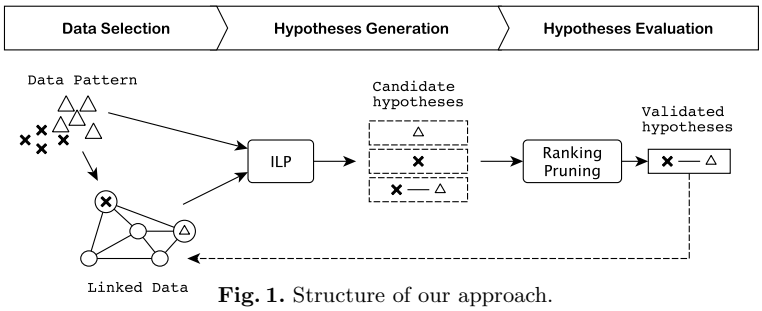


Fig. 1. Structure of our approach.

Data Selection. We introduced in the first section the RED example that we use to illustrate our approach. Once we obtain clusters of readers, we proceed with the search for information about them in Linked Data. For the purpose of a preliminary study, we started with the manual selection of some properties from DBpedia[6].

Hypotheses Generation. The step concerns the problem formulation in the ILP framework. Inductive Logic Programming is a research field at the inter-

Table 1. Prolog-encoded examples. Gordon Byron and Samuel Coleridge are examples of readers belonging to the same cluster `c`.

clusters	`c('Gordon Byron'). c('Samuel Coleridge').`
RDF predicates	`originCountry('Gordon Byron','England').`
RDF is-a relations	`country('England').`

section of Machine Learning and Logic Programming, investigating the inductive

[6] http://dbpedia.org/About

construction of first-order clausal theories (Logic Programming heritage) starting from a set of examples (Machine Learning heritage) [**?**]. Its distinguished feature is the use of some additional background knowledge to derive the hypotheses. In such framework, the data patterns represent the negative and positive examples, while information from Linked Data is the background knowledge required to generate hypotheses. Therefore, we encode them into Prolog clauses, as follows: The hypotheses are generated using the Aleph[7] system, and take the form of:

```
[Pos cover=14, Neg cover=308] c(A):-female(A)∧originCountry(A,'England')
```

which is interpreted as: "the reader `A` is part of the cluster `c` because of being `female` and from `England`". `Pos cover` is the number of examples e^+ covered by the rule r included in the cluster c ($e^+ \in c$), while `Neg cover` is the number of examples e^- covered by r, where $e^- \notin c$.

Hypotheses Evaluation. In this preliminary study, the hypotheses evaluation is performed using the weighted relative accuracy function (WR_{acc}) provided by Aleph and described in [**?**]. WR_{acc} measures the unusualness of a rule and expresses it in terms of number of positive and negative examples covered. By providing a trade off between of a rule's coverage and relative accuracy, WR_{acc} allows us to obtain explanations which are valid for patterns of small sizes. Given a rule r and a cluster c, WR_{acc} is defined as:

$$WR_{acc} = \frac{e_r^+ + e_r^-}{\mathcal{E}_c^+ + \mathcal{E}_c^-}\left(\frac{e_r^+}{e_r^+ + e_r^-} - \frac{\mathcal{E}_c^+}{\mathcal{E}_c^+ + \mathcal{E}_c^-}\right) \quad (1)$$

where e_r^+ and e_r^- the number of positive and negative examples covered by r, $\mathcal{E}^+$ the size of c and $\mathcal{E}^-$ the number of examples provided outside c. Using this formula, we obtained a preliminary ranking of the generated hypotheses. Examples of rules with the best scores are presented in Table **??**.

Table 2. Examples of generated hypothesis with their WR_{acc} score.

cluster	size	hypothesis	WR_{acc}
Austen J.	110	`c(A):- religion(A,'Anglican')`	0.025
		`c(A):- female(A)`	0.02
Pepys S.	13	`c(A):- religion(A,'Anglican')∧male(A)∧country(A,'England')`	0.025

7 Reflections

The previous table presents some promising results for the hypotheses evaluation, ranking and selection. The results for the first cluster are fairly strong when compared to the sample set ($\mathcal{E}^+ \cup \mathcal{E}^- = 1230$), and show how ILP is a good approach to explain data patterns, e.g. "*people reading Jane Austen were Anglican women*". This initial test also confirms our intuition that the proposed approach could naturally combine different sources of background knowledge (i.e., different datasets) to produce explanations of found patterns in the data. Here for example, information about the gender of readers come from the RED data, while the information about their religion is present in DBpedia. However,

[7] http://www.cs.ox.ac.uk/activities/machlearn/Aleph/

as expected, triggering a new background knowledge search process is required to make the explication more understandable. In practice, we might require a more specific answer to the question "what connects readers of Jane Austen?" than that they are Anglican women. We are also aware that finding a more adequate scoring measure to check the validity of a hypothesis is necessary. The WR_{acc} might be a good starting point but we will have to find an evaluation measure which takes into account aspects such as the lack of information or a smaller cluster size. This will, in fact, have a direct impact on the data selection. Finally, in order to detect what strongly connects the data in a pattern, we need to find a good way to detect valid background knowledge. Most of this PhD work will be focused on this issue.

8 Evaluation plan

(1) Hypothesis validity evaluation. We aim at finding the good rules using background knowledge from Linked Data. For instance, is "people reading Jane Austen were Anglican women" good, or good enough? Depending on the use-case we will be working on, a manual evaluation of the rules will be asked to the relevant domains experts.
(2) Experts support. How much our approach reduces the efforts needed from an expert? Does the explanation about the readers of Jane Austen bring any new knowledge to the expert, that he can exploit for the interpretation process? We will compare the results of a full KDD process achieved with and without our method to see whether the later can effectively reduce the expert's involvement.

9 Conclusions

This paper presents our research aiming at using background knowledge found in the Linked Data to explain patterns and regularities in data. The main idea is to explore if and how Linked Data can assist the experts in the knowledge discovery process. The first results of our ILP-based approach are promising and revealed that the Hypotheses Generation and Evaluation steps can be improved. We identified as one of the major issues the need of a **full access** to both data and the background knowledge. This information has to be (a) **expressive** (enough properties related to the data), **consistent** (no ambiguity or contradictory facts) and **complete** (properties need to cover most of the data). The future work will investigate the Data Selection step, the core part of our project. This PhD contribution will be to set up a good method to detect relevant information in Linked Data, where "detection" concerns both the right datasets and the represented data.

References

1. Antunes, C. (2008, October). An ontology-based framework for mining patterns in the presence of background knowledge. In *Int'l Conf. on Advanced Intelligence*, Beijing, China (pp. 163-168).
2. Brisson, L., Collard, M., & Pasquier, N. (2005, November). Improving the knowledge discovery process using ontologies. In Proceedings of the IEEE MCD international workshop on Mining Complex Data (pp. 25-32).

3. D'Aquin, M., Kronberger, G., & Suárez-Figueroa, M. (2012). Combining data mining and ontology engineering to enrich ontologies and linked data. In *Workshop: Knowledge Discovery and Data Mining Meets Linked Open Data-Know@ LOD* at Extended Semantic Web Conference, ESWC.
4. D'Aquin, M., & Jay, N. (2013). Interpreting Data Mining Results with Linked Data for Learning Analytics: Motivation, Case Study and Directions. In *Third Conference in Learning Analytics and Knowledge (LAK)*, Leuven, Belgium.
5. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
6. Grosof, B. N., Horrocks, I., Volz, R., & Decker, S. (2003, May). Description logic programs: Combining logic programs with description logic. In *Proceedings of the 12th international conference on World Wide Web* (pp. 48-57). ACM.
7. Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, 1(1), 1-136.
8. Hartmann, J., & Sure, Y. (2004, July). A knowledge discovery workbench for the Semantic Web. In International Workshop on Mining for and from the Semantic Web (p. 56).
9. Lisi, F. A. (2010). Inductive Logic Programming in Databases: From Datalog to DL+log. *Theory and Practice of Logic Programming*, 10(03), 331-359.
10. Liu, J., Wang, W., & Yang, J. (2004, August). A framework for ontology-driven subspace clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 623-628). ACM.
11. Marinica, C., & Guillet, F. (2010). Knowledge-based interactive postmining of association rules using ontologies. *Knowledge and Data Engineering*, IEEE Transactions on, 22(6), 784-797.
12. Moss, L., Sleeman, D., Sim, M., Booth, M., Daniel, M., Donaldson, L., & Kinsella, J. (2010). Ontology-driven hypothesis generation to explain anomalous patient responses to treatment. *Knowledge-Based Systems*, 23(4), 309-315.
13. Motik, B., & Rosati, R. (2006). Closing semantic web ontologies. Technical report, University of Manchester, UK.
14. Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19, 629-679.
15. Narasimha, V., Kappara, P., Ichise, R., & Vyas, O. P. (2011). LiDDM: A Data Mining System for Linked Data.
16. Novak, P. K., Vavpetic, A., Trajkovski, I., & Lavrac, N. (2009). Towards semantic data mining with g-segs. In Proceedings of the 11th International Multiconference Information Society, IS.
17. Paulheim, H., & Fümkranz, J. (2012, June). Unsupervised generation of data mining features from linked open data. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (p. 31). ACM.
18. Pan, D., Shen, J. Y., & Zhou, M. X. (2006). Incorporating domain knowledge into data mining process: An ontology based framework. *Wuhan University Journal of Natural Sciences*, 11(1), 165-169.
19. Phillips, J., & Buchanan, B. G. (2001, October). Ontology-guided knowledge discovery in databases. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 123-130). ACM.
20. Verborgh, R., Van Deursen, D., Mannens, E., & Van de Walle, R. (2010). Enabling advanced context-based multimedia interpretation using linked data.

Adaptive Navigation through Semantic Annotations and Service Descriptions

Ruben Verborgh

supervised by
Rik Van de Walle

Ghent University – iMinds – Multimedia Lab
Gaston Crommenlaan 8 bus 201, 9050 Ghent, Belgium
ruben.verborgh@ugent.be

Abstract. While hyperlinks are absolutely crucial to the Web's success, they are currently uni-directional, as information is augmented with controls from the perspective of the information publisher. However, it is the user who needs those links to navigate—and the publisher cannot know how any user might want to interact with the information. Therefore, the most relevant links for a user might be omitted, severely limiting the applicability of hypertext. In this paper, I outline a plan to tackle this problem as part of my doctoral research, by explaining the research questions, the underlying hypotheses, and the approach, in which semantic technologies play a crucial role.

Keywords: affordance, hypermedia, Semantic Web, Web services

1 Problem Statement

The hyperlink-driven information model of the World Wide Web [3] has introduced humanity to a novel way of information consumption. Information has become *actionable* [11], in contrast to the passive medium it used to be. Although hypertext was envisioned long before [15], the Web was the first hypertext system that worked on a global scale. Still, the initial visions of hypermedia featured a much richer arsenal of link mechanisms, such as non-breaking n-way links and transclusion [16].

The main problem of hyperlinks on the Web is that the responsibility of link creation lies entirely with the publisher. Indeed, when creating a Web site or application, it is the publisher of the information who decides what actions the consumers of that information can perform. This poses a threat to the said actionability of the information, because it can only be called "actionable" to the extent the consumer can actually perform the actions he or she desires. If the publisher omits the hyperlinks that afford these actions, then the hypertext document becomes as passive as any pre-Web medium, defeating its purpose.

The problem statement of my doctoral research is therefore how we can enhance the controls in hypermedia documents on the Web in a personalized way,

such that they afford the relevant actions for each consumer. I want to look at this problem from the standpoint of both human visitors of websites and machine clients of Web APIs, as they each have unique challenges. The aim is to provide these controls in an automated way with the least possible amount of human intervention.

In the next section, I will explain the relevancy of this problem, followed by a summary of related work in Section 3. Section 4 poses several research questions, for which I formulate hypotheses in Section 5. My approach is presented in Section 6, followed by a reflection in Section 7. Section 8 discusses the evaluation and preliminary results are shown in Section 9.

2 Relevancy

It is crucial to realize that hyperlinks—and hypermedia controls in general—are not *enablers* but *affordances* [11,17], *i.e.*, they offer the information consumer an action possibility, but the action itself is also achievable through other means. For instance, if a document refers to another page without a hyperlink, that page might still be accessible (*e.g.*, through a search engine). However, this catapults us back to the age of paper documents, as the document does not directly contain the needed affordances, defeating the purpose of hypertext. Furthermore, such indirect ways are more time- and resource-consuming, while the ability to browse quickly in a hypermedia system is critical to its usability [2]. Thus, having the right hypermedia controls in place is necessary for efficient Web use. Three reasons in particular make the need for more relevant controls very actual.

Continued expansion The Web is growing at an ever increasing rate, which means that if the average number of links per page does not increase, the connectedness of the Web is decreasing. In 1999, the size of the Web was estimated at 800 million documents and the average document distance already at 19 clicks [19]—and it is not hard to imagine that this distance will only go up. This means that the trade-off between *completeness* (a publisher offering all relevant links) and *efficiency* (a consumer quickly finding the link she needs) becomes increasingly difficult to manage.

Aging documents Even if publishers could somehow strike an ideal balance between completeness and efficiency at the time of publication, it is highly unlikely that their choice of links will remain optimal as the document ages. Since the Web does not implement the concept of backlinks, the fact that new documents can link to older ones does not improve the affordance of the older documents. Furthermore, links to certain pages break if documents move or cease to exist [14]. Therefore, the hypermedia controls found on older documents are not the most relevant ones for a user. For instance, given the current trend of online social networks, many recent sites provide interaction controls with those networks. When a user browses older pages, these controls can be missed, especially by users who lack the necessary skills to perform these actions without the help of such designated controls.

Mobile growth In the past few years, mobile use of the Web has increased tremendously and will continue to do so in the coming years [10]. The nature of mobile devices makes the Web browsing experience different: average screen sizes are smaller and physical keyboards are either miniaturized or not present. The small screen size leaves less space for links, so the few that appear must be relevant. Also, the absence of a full-scale keyboard makes it more difficult to reach a goal in indirect ways (such as entering keywords in a search engine) if a direct link that leads to the user's desired goal is missing.

3 Related Work

3.1 Adaptive Hypermedia

The personalization of hypermedia documents is part of the research field of *adaptive hypermedia* [6]. Within adaptive hypermedia research, *adaptive navigation support* [8] is concerned with personalizing hypermedia controls to match the intentions or goals of the user. Most adaptive navigation support systems are *a)* operating on a closed corpus, *b)* focused on linking to related information, and *c)* used in a specific context such as learning. In contrast, we want to approach the problem statement *a)* on the open corpus of the entire Web, *b)* with a focus on performing actions, and *c)* for day-to-day usage. In fact, *open-corpus adaptive hypermedia* has been identified as an important challenge in the field [7], but it has not been tackled intensively. Semantic Web technologies were listed as a candidate to help overcome the open-corpus problem on the Web [9].

3.2 Social Interaction Widgets

In response to the sudden rise of many social networks, publishers started adding so-called *widgets* to their sites, small snippets of code that provide user interactions. In contrast with simple hyperlinks, which connect one document to another, these widgets typically allow the user to perform an *action* on the current document, such as sharing it within a social circle or adding a comment. Examples include the Facebook *Like* button and the Twitter *Tweet* button [5]. However, as more social networks emerge, it becomes impractical for publishers to manually add widgets for each of them.

In order to cope with this increasing diversity in social networks, aggregated widgets such as AddThis [1] were created. AddThis is a single widget that gives access to sharing options on various social networks through a list that every user can personalize. The benefit on the publisher side is that he does not need to know about the user's preferred network, nor must provide a sharing link. Additionally, the user is not bothered by non-relevant sharing links, because social networks that she does not use are not needlessly displayed by the widget. However, the usage of AddThis is limited to sharing actions, and it is thereby not a generalized solution for personalized action links on the Web.

3.3 Web Intents

A solution that does support a wider range of actions is Web Intents [4]. The idea derives from a concept on phones with the Android operating system, where applications can indicate their *intent* to support a certain action, such as calling or sending a message. The specification defines several standard actions, such as sharing, editing, and viewing, which can be supported by Web applications. Content publishers should indicate which actions their users can perform. However, this still implies the publisher must "predict" what type of action the user might want to do [21]. In contrast, we want to determine the action through the user's preferences, which are highly personal and can change over time.

4 Research Questions

The main question in my doctoral research is:

> *How can we enhance hypermedia controls in a personalized way, such that they complement a piece of information with the affordance a user requires to perform the next steps he or she needs?*

This question gives rise to two others. On the one hand, there is the human aspect:

> *How does such enhanced affordance help users browsing the Web in achieving their goals, and how can we achieve maximum effectiveness in this regard?*

And on the other hand, there is the machine aspect:

> *How does enhanced affordance help machine clients consume Web APIs, and can it lead to true serendipitous reuse [23] of services?*

This last question is inspired by the concept *hypermedia as the engine of application state* [12], which aims to achieve loose conversational coupling [18] by augmenting representations with controls, also for machine clients. However, as is the case with humans surfing the Web, publishers of information are unaware of the goals of the information consumer, and it is therefore hard for them to provide the necessary affordance [21].

5 Hypotheses

The main hypothesis, related to the main research question, is:

> *Current semantic technologies are sufficiently flexible and powerful to connect a piece of information and matching actions at runtime, while providing loose coupling at design time.*

In addition to the feasibility, the following hypothesis relates to the necessity:

> *Semantic technologies offer an added value to the creation process of enhanced affordance.*

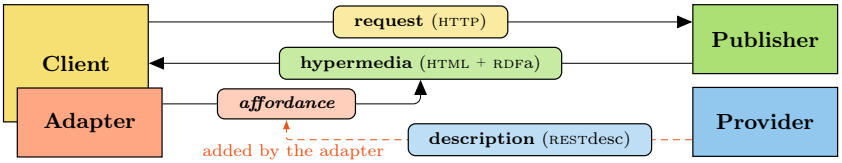


Fig. 1. An adapter at the *client* side adds affordance to the hypermedia representation, based on the semantic annotations the latter contains (*e.g.*, RDFa or HTML5 microdata).

My hypotheses regarding the effectiveness of hypermedia documents enhanced with personalized affordance are:

> *Users can browse the Web faster and more efficient when the relevant affordance is in place.*
>
> *Machine clients will be more functional and more resilient to change if they receive messages with enhanced affordance.*

6 Approach

My approach to address the research questions is to develop a technology and architecture for what I call *distributed affordance* [21]. The core idea is that publishers add *semantic annotations* to the hypermedia documents they serve to a client, which are matched at runtime by the client to *semantic service descriptions* that describe the *functionality* offered by providers of the user's choice. Figure 1 shows a client making a request to an information publisher, who replies with an HTML document that has been enriched with RDFa markup. Earlier, a client-side adapter has accessed functional descriptions in RESTdesc [22] format, which are now instantiated with the RDFa annotations to generate affordances that can be added to the HTML document. These affordances will lead to actions that operate directly on the resources inside the page.

This addresses the main research question, and will also put the main hypothesis to the test. The proposed platform is loosely coupled, as the publisher, provider, and adapter do not need to know about each other. Instead, the publisher and provider offer sufficient semantics for the adapter to interpret what combinations are possible. This contrasts with current hyperlinks on the one hand, where the publisher has to know about the action provider, and with closed-corpus adaptation on the other hand, where the adapter has to know about the publisher and/or the action provider.

To address the research questions and hypothesis about user and machine usage of the distributed affordance platform, I will create a corpus consisting of websites with annotations and Web services with descriptions. The goal is to incorporate as many real-world examples as possible, in order to have a realistic testing environment, as well as several use cases wherein the technology proves its added value.

7 Reflections

The main difference in my approach with existing work on adaptive navigation support [8], is that I specifically want to perform adaptation on an *open* corpus, *i.e.,* the entire Web, instead of a controlled subset. Open corpus adaptive hypermedia has in fact been identified as an important challenge [7], and semantic technologies have been hinted at as a potential solution [9], although no concrete systems have emerged yet. However, I believe that the Web only recently is becoming ready for this, as it is only in the past few years that we see sufficient semantic annotations appear, despite the technologies (such as RDFa) being around for a longer time. Furthermore, my previous experimentation with functional description of REST APIs [22] gives me the confidence that this technology is sufficiently mature to apply it on automated action creation using even limited semantic annotations.

A second difference, as indicated in Figure 1, is that the adaptation happens at the *client* side and therefore is fully scalable, instead of classical adaptation systems that rely on a central adaptation component. Furthermore, whereas the majority of work on adaptive navigation focuses on linking static documents together, my goal is to connect *dynamic* documents, *i.e.,* generate links towards actions on the current document. This allows for the creation of much richer interactions.

8 Evaluation Plan

For the evaluation, there are three main lines of interest.

user studies As the main focus is on creating affordance that will help users browse more efficiently, it is of utmost importance to conduct user studies that follow people's browsing behavior as they make use of the developed platform. Several tests should be conducted in a (double-)blind setting, where the participant (and possibly the experiment conductor) are unaware whether the platform is active. I will analyze qualitative parameters on the one hand, such as the user's impression of effectiveness, and quantitative parameters on the other hand, such as the time to complete a task.

performance evaluation In a platform that should manipulate web pages in real-time, speed will be crucial. Therefore, various aspects of the platform should be tested for performance, especially the semantic matching of content and services, which can become complex quickly. In addition to that, the whole pipeline must be tested, and optimized so that it stays under the threshold that is deemed acceptable in the user studies.

client code complexity Finally, as I also want to focus on automated clients of Web APIs, the code of such clients should be less complex [13] as a result of the enhanced affordance. This makes it necessary to compare the implementations of clients with and without the use of (additional) hypermedia controls in the server's response.

9 Preliminary Results

Previously, I have evaluated the performance of Web API matching and composition with RESTdesc [20], which led me to conclude two things. First, it is possible to create relevant composition chains in a few milliseconds. Second, this level of performance can be maintained even with thousands of different descriptions of Web site actions that are potential matches. Together, this indicates that finding the few API descriptions that match a given resource out of a large description set is possible in a reasonable amount of time.

Additionally, I have started a user study in collaboration with researchers from the human-computer interaction field, in which we observe users as they perform tasks on the Web with and without the distributed affordance platform. The first results seem to suggest that users navigate indeed more efficiently when the affordance has been optimized for their needs. The distributed affordance platform itself is currently under development, the progress of which can be followed at `http://distributedaffordance.org/`.

10 Conclusion

This paper has presented the outline of my doctoral research, in which I want to focus on personalized affordance created from distributed sources. My approach is to build a platform that works on the client side, enhancing hypermedia representations returned by the server with hyperlinks and controls of the user's preference. Semantic technologies enable a loose coupling between publishers of information and action providers, which allows the platform to have a truly distributed nature. The evaluation of this work will consist of user studies, performance evaluations and code complexity comparisons. My goals are to make browsing the Web more efficient for users, and to enable a more serendipitous reuse of services for machines.

Acknowledgements The described research activities were funded by Ghent University, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union.

References

1. AddThis, `http://www.addthis.com/`
2. Akscyn, R.M., McCracken, D.L., Yoder, E.A.: KMS: a distributed hypermedia system for managing knowledge in organizations. Communications of the ACM 31(7), 820–835 (Jul 1988)
3. Berners-Lee, T., Cailliau, R., Groff, J.F.: The world-wide web. Computer Networks and ISDN Systems 25(4–5), 454–459 (1992), `http://www.sciencedirect.com/science/article/pii/016975529290039S`
4. Billock, G., Hawkins, J., Kinlan, P.: Web Intents. W3C working draft (Jun 2012), `http://www.w3.org/TR/web-intents/`
5. Bodle, R.: Regimes of sharing. Information, Communication and Society 14(3), 320–337 (Apr 2011)

6. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction 6(2–3), 87–129 (1996), http://dx.doi.org/10.1007/BF00143964
7. Brusilovsky, P.: Adaptive hypermedia. User Modeling and User-Adapted Interaction 11(1–2), 87–110 (2001), http://dx.doi.org/10.1023/A%3A1011143116306
8. Brusilovsky, P.: Adaptive navigation support. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, pp. 263–290. Springer-Verlag (2007), http://dl.acm.org/citation.cfm?id=1768197.1768207
9. Brusilovsky, P., Henze, N.: Open corpus adaptive educational hypermedia. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, pp. 671–696. Springer-Verlag, Berlin, Heidelberg (2007), http://dl.acm.org/citation.cfm?id=1768197.1768223
10. Cisco: Visual networking index: Global mobile data traffic forecast update, 2012–2017 (Feb 2013), http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf
11. Fielding, R.T.: REST APIs must be hypertext-driven. Untangled – Musings of Roy T. Fielding (Oct 2008), http://roy.gbiv.com/untangled/2008/rest-apis-must-be-hypertext-driven
12. Fielding, R.T., Taylor, R.N.: Principled design of the modern Web architecture. Transactions on Internet Technology 2(2), 115–150 (May 2002)
13. Lanning, D., Khoshgoftaar, T.: Modeling the relationship between source code complexity and maintenance difficulty. Computer 27(9), 35–40 (1994)
14. Markwell, J., Brooks, D.W.: Broken links: the ephemeral nature of educational www hyperlinks. Journal of Science Education and Technology 11(2), 105–108 (2002)
15. Nelson, T.: Complex information processing: a file structure for the complex, the changing and the indeterminate. In: Proceedings of the ACM 20th National Conference. pp. 84–100. ACM, New York, NY, USA (1965)
16. Nelson, T.: Dream machines. self-published (1974)
17. Norman, D.A.: The Design of Everyday Things. Doubleday, New York (1988)
18. Pautasso, C., Wilde, E.: Why is the Web loosely coupled? – A multi-faceted metric for service design. In: Proceedings of the 18th international conference on World Wide Web. pp. 911–920. ACM, New York (2009), http://doi.acm.org/10.1145/1526709.1526832
19. Reka, A., Jeong, H., Barabasi, A.L.: Diameter of the World-Wide Web. Nature 401(6749), 130–131
20. Verborgh, R., Haerinck, V., Steiner, T., Van Deursen, D., Van Hoecke, S., De Roo, J., Van de Walle, R., Gabarró Vallés, J.: Functional composition of sensor Web APIs. In: Proceedings of the 5th International Workshop on Semantic Sensor Networks (Nov 2012), http://ceur-ws.org/Vol-904/paper6.pdf
21. Verborgh, R., Hausenblas, M., Steiner, T., Mannens, E., Van de Walle, R.: Distributed affordance: An open-world assumption for hypermedia. In: Proceedings of the Fourth International Workshop on RESTful Design (May 2013), http://distributedaffordance.org/publications/ws-rest2013.pdf
22. Verborgh, R., Steiner, T., Van Deursen, D., Coppens, S., Gabarró Vallés, J., Van de Walle, R.: Functional descriptions as the bridge between hypermedia APIs and the Semantic Web. In: Proceedings of the Third International Workshop on RESTful Design. pp. 33–40. ACM (Apr 2012), http://www.ws-rest.org/2012/proc/a5-9-verborgh.pdf
23. Vinoski, S.: Serendipitous reuse. Internet Computing 12(1), 84–87 (2008), http://steve.vinoski.net/pdf/IEEE-Serendipitous_Reuse.pdf