

Pattern discovery from the ontological layer of the Semantic Web (KWEPSY2006 – extended abstract)

Agnieszka Ławrynowicz

Institute of Computing Science, Poznan University of Technology
ul. Piotrowo 2, 60-965, Poznan, Poland
agnieszka.lawrynowicz@cs.put.poznan.pl
<http://www.cs.put.poznan.pl/alawrynowicz>

Abstract. The topic of my PhD lies within Semantic Web Mining. It concerns knowledge discovery from rich structures, expressed in the languages from the ontological layer of the Semantic Web. It aims at developing data mining techniques that use ontologies as background/domain knowledge. In particular, background knowledge in this approach can be expressed in hybrid languages that can be seen as subsets of SWRL, lying on the intersection/combination of OWL DL and Logic Programming (e.g. OWL DLP or DL-safe rules). I am interested in investigating the properties of these, hybrid languages for the tasks of knowledge discovery. For doctoral research I have chosen one task – frequent pattern discovery.

1. Problem description

The problem of knowledge discovery from datasets expressed in the Semantic Web languages can be potentially important when there is a lot of scientific and other, highly formalized, data published on the Web. This data can be then queried and searched. Across these huge datasets, hypotheses can be generated and verified (for example from life sciences domain) thus leading to derivation of new knowledge from existing data. It would be very beneficial for the Semantic Web users if new and interesting knowledge could be discovered from the existing datasets in an automatic way. Thus the application of mining huge, formalized, Web repositories seems promising and interesting.

Data mining

Discovery of new and potentially interesting knowledge from large data sets is the task of *data mining*. Data mining methods are developed to search for *patterns* in *data*. Patterns can be discovered from unstructured data, semi-structured documents or from structured sources like relational databases. Most of the methods of data mining proposed so far operate however on unstructured data (in a form of a single table with data). This “attribute-value” representation requires the data to be preprocessed and aggregated into single table risking loss of meaning or information. In recent years, however, the methods operating on original, non-preprocessed data sets have received more attention. In the structure of the data implicit semantics is hidden that can be exploited by such methods. Semantic information can also be included explicitly in data mining process in the form of rules and relationships existing in given domain (so called *background/prior knowledge*). To the methods that mine patterns in structured datasets belong *relational data mining*, *RDM* approaches that mine patterns from relational data bases. RDM approaches are part of a larger group of methods of *inductive logic programming*, *ILP*. In the case of ILP methods the data is represented in logic-based languages like for example Datalog. In the Semantic Web knowledge bases can be also represented in languages based on logic-based formalisms like *OWL* which is based on *description logics*, *DL*. The goal of my doctoral research is to provide sound methods for mining data sets expressed in the languages from the ontological layer of the Semantic Web, which can be referred as to *Semantic Web Mining*.

Semantic Web mining

Semantic Web Mining can be viewed from different angles. From one side, it can be seen as using the data/Web mining techniques to build ontologies (“mining to”). From the other side it can be seen as how to use the RDF, ontologies, logics to support the process of data/Web mining (“mining from”). The first approach is now widely recognised as valuable in the process of ontology learning (by knowledge discovery from textual, semi-structured resources). The latter one has not been yet intensively studied. Recently, however, there is increasing interest in using explicitly and formally represented prior knowledge. Such information may help in the selection of data, pruning the hypothesis space and in better understanding of obtained results. Formal specification of background knowledge in the form of an ontology may help in efficient automatization of knowledge discovery process. Obtained results may be then automatically published on the Semantic Web. Domain ontologies may also help in handling efficiently complex objects during mining the data for Semantic Web personalisation.

2. Related work

The work related to mine can be divided into two groups: relational data mining and data mining from more expressive representations.

Relational data mining

The problem of discovery of frequent relational patterns was introduced by [1]. As a solution to this problem an ILP method called WARMR was proposed. WARMR is an adaptation of the levelwise method, originally used in APRIORI algorithm operating on item sets. WARMR instead of item sets operates on atom sets (conjunctive queries in Datalog). In WARMR, the space of patterns is searched one level at a time starting from the most general patterns and iterating between candidate generation and candidate evaluation phases. As a generality measure WARMR uses approximation of logical implication called θ -subsumption. WARMR performs a lot of tests for equivalence under θ -subsumption in order to prune infrequent and redundant queries during pattern generation phase. An early version of WARMR was inefficient, thus it has been further optimized in many different ways. In [9] another RDM method for frequent pattern discovery named FARMER was introduced. FARMER also uses the notation of first order logic, but it does not depend on a time consuming test for equivalence. In FARMER the special data structure called trie, inspired by the implementation of APRIORI, is used instead. Under some restrictions on the language FARMER is equivalent to WARMR and achieves better performance.

Data mining from more expressive representations

Although relational data mining methods are proved to be useful they have also some drawbacks. Firstly, θ -subsumption is not fully semantic measure as it is not equal to logical implication which results in methods that are incomplete. It was chosen because computing the logical implication in this case is undecidable. Horn rules as a representation language are also not very well suited for modeling hierarchical structures. Description logic, in turn, was developed to be able to represent rich structural knowledge. From the other side, description logic does not allow for the interaction of variables in arbitrary ways, which is in turn the property of Horn clausal logic. To benefit from strong parts of both formalisms, the combination of expressive power of DL and Horn clausal logic as a representation language in data mining seems to be desirable. To the best of my knowledge, there is only one approach, named SPADA [6], that is developed to use such an expressive representation for frequent pattern mining. SPADA uses hybrid \mathcal{AL} -log language. The current version of

SPADA admits however very basic language in DL component. Also the task of frequent pattern discovery is formulated in such a way that either the patterns that can be found contain concepts only from the same level of taxonomy or some concepts are replicated in some, lower levels of taxonomy.

3. Proposed approach

In my approach to pattern mining I have decided to use recently introduced combination of DL and function free Horn rules, so-called \mathcal{DL} -safe rules [7]. This combination allows using very expressive DL, while still preserving the decidability property. Recent tests [8] show also that KAON2, reasoner implementing this approach, outperforms another reasoners in case of high number of instances in knowledge base which is exactly the case in my research problem. As a starting point of my research on mining from the Semantic Web I have decided to use the simplest language – OWL DLP.

Data mining tasks are defined by specification of kind of patterns searched, the data in which the patterns are mined (*extensional* background knowledge, instances) and *intensional* background knowledge in the form of general rules describing given domain. In my approach I assume pattern mining in knowledge bases KB represented in OWL DLP, which contain the terminological (TBox) and the assertional (ABox) parts consistent with each other. The intensional background knowledge is represented in TBox. The extensional background knowledge (instances) is represented in ABox. The goal is to find frequent patterns in the form of conjunctive queries over KB. I have formulated the task of knowledge discovery for this setting as follows:

Definition 1. Given

- a knowledge base in OWL-DLP with \mathcal{DL} -safe rules \mathcal{KB} ,
- a set of patterns in the language \mathcal{L} of queries Q that all contain a reference concept C_{ref} ,
- a minimum support threshold *minsup* specified by the user

and assuming that queries with support s are frequent in \mathcal{KB} given C_{ref} , denoted as $support(C_{ref}, Q, \mathcal{KB})$, if $s \geq minsup$, the task of *frequent pattern discovery* is to find the set \mathcal{F} of frequent queries.

Queries are conjunctive \mathcal{DL} -safe queries. The C_{ref} parameter determines what is counted. The atom of the query that has as a predicate the reference concept, contains the only one distinguished variable that can appear in the query (*key* variable).

Definition 2. A *support* of the query Q with respect to the knowledge base \mathcal{KB} is defined as the ratio between the number of instances of the C_{ref} concept that satisfy the query Q and the total number of instances of the C_{ref} concept (obtained as a result of submitting a trivial query denoted Q_{ref}):

$$support(C_{ref}, Q, \mathcal{KB}) = \frac{|answerset(C_{ref}, Q, \mathcal{KB})|}{|answerset(C_{ref}, Q_{ref}, \mathcal{KB})|}$$

Here is the example of the query: $q(key):-Client(key), isOwnerOf(key, x), Account(x)$.

In [3] this new data mining setting and its potential has been introduced by my research group. A method inspired by early version of WARMR has been presented together with a case study on financial ontology. Pattern mining in this method took considerable amount of time, because candidate pattern generation mechanism barely benefited from what was found in previous levels. As this method was not optimized it can be seen as a proof-of-concept.

Next, in [4] the method based on a special trie structure, similar to that used in APRIORI and FARMER has been presented, where considerable speedup as compared to early, naive approach has been obtained. In both cases OWL DLP language was taken into consideration, but in the latter case the language was further significantly restricted. In current work [5] I develop the algorithms and optimization techniques for pattern discovery from knowledge bases in unrestricted OWL DLP language. My current algorithms are based on the idea of levelwise search, where the space of patterns is searched one level at a time starting from the most general patterns. The algorithms use a trie data structure in which the nodes correspond to the atoms of the query and every path from the root to a node corresponds to a query. New nodes are added to the trie only to the leaves that correspond to frequent queries. Following the classification introduced in [9] three ways in which atoms can be added as leaves to the trie are distinguished:

Definition 3. Refinement rules. Atoms are added to the trie as:

1. *dependent atoms* (which use at least one variable of the last atom in the query),
2. *right brothers* of a given node,
3. *a copy* of a given node.

For each predicate a list of admissible predicates from which dependent atoms of this predicate can be built is computed. The way in which the predicates can be used in dependent atoms is also determined (which variables can be shared with the parent node). These information is computed according to the intensional part of background knowledge in TBox. Dependent atoms are added as children of a given node as well as are its right brothers in the trie. Right brother copying mechanism takes care that all possible subsets are generated and (thanks to the trie structure) only one permutation out of a set of dependent atoms is considered. If necessary, variable names of right brothers are changed when they are being copied. Every time when new node is being added semantic information from TBox is consulted. In Figure 1 an example trie is presented (the numbers on edges refer to three ways in which atoms can be added to a trie):

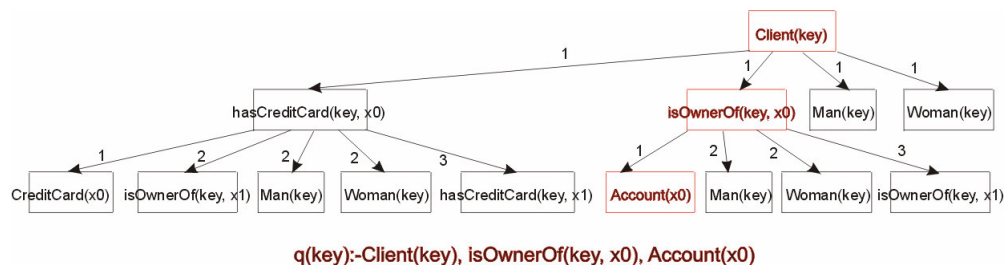


Figure 1. The part of the trie generated when asking about *Client* as a reference concept, for a simple ontology with the concepts *Client* (subconcepts: *Man*, *Woman*), *CreditCard* and *Account* and with the properties *hasCreditCard* and *isOwnerOf*.

In further work on the task of frequent pattern mining I would like to investigate more expressive languages from \mathcal{DL} -safe rules family.

In my approach, I take into account hybrid language, combining description logic and Horn clausal logic. As it has been mentioned there exists only one other approach that aims at frequent pattern discovery using such an expressive language, namely system SPADA. In SPADA, Horn rules component is unrestricted, but description logic component is quite restricted. I plan in turn to investigate deeper in my approach the description logic component. In the case of OWL DLP, I also apply fully semantic generality measure based on query containment (not θ -subsumption).

4. Expected contributions of the approach

An expected contribution of my work is the development of the new class of data mining methods that mine patterns from expressive, hybrid languages from the ontological layer of the Semantic Web. In most of the current knowledge discovery methods, the background knowledge is implicit or does not have formal structure and semantics. It can be then practically considered only by the human analyst. Such approaches does not take into account the progress in the field of knowledge engineering where domain knowledge can be represented formally by ontologies. The approaches where ontologies and data mining methods are combined to discover, interpret and possibly reorganise knowledge has not been studied extensively yet, and my approach is one of the very first ones.

To the benefits of this approach can belong also the investigation of how to efficiently traverse the space of patterns. The results of such an investigation can then be further used in another data mining tasks. In my approach, ontologies may not only be taken as an input, but discovered knowledge may serve in turn for ontology evaluation or evolution. Discovered patterns can be processed further into association rules or used for conceptual clustering. Each query can represent a cluster of individuals of the reference concept and cluster hierarchy can be built according to query containment between queries.

5. Research methodology

For every presented algorithm I assume to prove that it is complete (generates every possible pattern from given language of patterns). I plan to make throughout performance evaluation of my algorithms. In particular I plan to do the tests investigating the potential benefits of guiding the search for patterns by ontology (terminological part) when compared to the naive approach where every possible pattern is generated. I am going to investigate different sizes and complexities of terminological parts as well as of assertional parts of knowledge bases. I also plan to develop heuristics.

More on my doctoral research can be found on SEMINTEC¹ project site.

References

1. Dehaspe, L., Toivonen, H.: Discovery of frequent Datalog patterns. *Data Mining and Knowledge Discovery*, 3(1): 7 - 36, (1999)
2. Grosz B. N., Horrocks I., Volz R., and S. Decker. Description Logic Programs: Combining Logic Programs with Description Logic. In *Proc. of the Twelfth Int'l World Wide Web Conf. (WWW 2003)*, pages 48–57. ACM, (2003)
3. Józefowska J., Ławrynowicz A., Łukaszewski T. (2005) Towards discovery of frequent patterns in description logics with rules, *Proc. of the International Conference on Rules and Rule Markup Languages for the Semantic Web (RuleML-2005)*, Galway, Ireland, LNCS, Springer-Verlag, 84-97
4. Józefowska J., Ławrynowicz A., Łukaszewski T. Faster frequent pattern mining from the Semantic Web, *IIS: IIPWM'06, Advances in Soft Computing*, Springer Verlag 2006, accepted for publication
5. Józefowska J., Ławrynowicz A., Łukaszewski T., Frequent pattern discovery from OWL DLP knowledge bases, Submitted for publication
6. Lisi F.A., Malerba D., Inducing Multi-Level Association Rules from Multiple Relation, *Machine Learning Journal*, 55, 175-210, (2004)
7. Motik B., Sattler U., Studer R.. Query Answering for OWL-DL with Rules. *Proc. of the 3rd International Semantic Web Conference (ISWC 2004)*, Hiroshima, Japan, November, 2004, pp. 549-563, (2004)
8. Motik B., Sattler U. Practical DL Reasoning over Large ABoxes with KAON2. Submitted for publication
9. Nijssen, S., Kok, J.N. (2001) Faster Association Rules for Multiple Relations. *Proceedings of the IJCAI'01*, 891-897

¹ <http://www.cs.put.poznan.pl/alawrynowicz/semintec.htm>