Knowledge-Based Systems 25 (2012) 51–62

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Finding association rules in semantic web data

Victoria Nebot *, Rafael Berlanga

Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Campus de Riu Sec, 12071 Castellón, Spain

ARTICLE INFO

Article history:
Available online 26 May 2011

Keywords:
Semantic web
Data mining
Association rules
Semantic annotation
Biomedical application

ABSTRACT

The amount of ontologies and semantic annotations available on the Web is constantly growing. This new type of complex and heterogeneous graph-structured data raises new challenges for the data mining community. In this paper, we present a novel method for mining association rules from semantic instance data repositories expressed in RDF/(S) and OWL. We take advantage of the schema-level (i.e. *Tbox*) knowledge encoded in the ontology to derive appropriate transactions which will later feed traditional association rules algorithms. This process is guided by the analyst requirements, expressed in the form of query patterns. Initial experiments performed on semantic data of a biomedical application show the usefulness and efficiency of the approach.

1. Introduction

In the past few years, there has been an increasing interest in combining the two research areas semantic web (SW) and data mining (DM) [1,33]. Thanks to the standardization of the ontology languages RDF/(S)[1] and OWL,[2] the SW has been realized and the amount of available semantic annotations is ever increasing. This is due in part to the active research concerned about learning knowledge structures from textual data, usually referred as Ontology Learning [9]. On the other hand, background knowledge has been also used to improve the results of Web mining. However, little work has been directed towards mining SW data itself, that is, mining the SW. We strongly believe that mining SW data will bring much benefit to many domain-specific research communities where relevant data are often complex and heterogeneous, and a large body of knowledge is available in the form of ontologies and semantic annotations. This is the case of the clinical and biomedical scenarios, where applications often have to deal with large volumes of complex data sets with different structure and semantics. In this paper, we investigate how ontological instances expressed in OWL can be combined into transactions in order to be processed by traditional association rules algorithms [2], and how we can exploit the rich knowledge encoded in the respective ontologies to reduce the search space.

Machine learning algorithms have been successfully applied to large data sets to extract useful knowledge by searching for interesting patterns (e.g., association rules). However, the nature of semantic data is quite different from that of traditional data sets treated by these algorithms. Thus, the main challenges we face in this work are the following ones:

- Traditional DM algorithms deal with homogeneous data sets composed by transactions, where each transaction is represented by a subset of items. In contrast, in a repository of semantic annotations expressed in OWL we keep ontology axioms, describing the conceptual domain, and the semantic annotations are represented as assertions relating instances through properties that are consistent with the ontology. The usual way to represent these assertions is as triples (*subject*,*predicate*,*object*). In this scenario, the identification of transactions and items is not trivial. Items may correspond to either instances or literals, and a transaction is defined according to the user requirements as a subset of items semantically related in the repository.
- OWL is sustained by description logics (DLs) [6], which are knowledge representation formalisms with well-understood formal properties and semantics. Therefore, annotated data does not follow a rigid structure. That is, instances belonging to the same OWL class may have different structures, giving place to structural heterogeneity issues.
- Since DLs are defined with formal semantics, reasoning capabilities must be applied in order to handle the implicit knowledge.

As far as we know, the presented approach is the first attempt to find association rules directly from SW data. Previous work on mining SW data has been mainly focused on clustering and classifying ontology instances [8,12]. However, the representation techniques required in association mining are quite different from clustering and classification tasks. Association rules are based on

* Corresponding author. Tel.: +34 964 72 83 67; fax: +34 964 72 84 35.
E-mail addresses: romerom@lsi.uji.es (V. Nebot), berlanga@lsi.uji.es (R. Berlanga).

[1] RDF/(S): http://www.w3.org/TR/rdf-concepts/,rdf-schema/, '04.
[2] OWL: http://www.w3.org/TR/owl-features/, '04.

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.05.009

the notion of transaction, which is an observation of the co-occurrence of a set of items. This is basically a set-based representation of the world which contrasts with the numerical vector-based representations used in clustering and classification. When dealing with SW data, the main challenge consists in identifying interesting transactions and items from the semi-structured and heterogeneous nature of these data. In this way, it becomes crucial to use as much as possible the knowledge provided by the ontologies so that transactions can be easily defined and mined. As we will show along this paper, there is a great variety of ways to generate items and transactions from semantic data, which depend on the detail level and structural semantics the analyst wants to consider.

The main contribution of this paper is twofold: (1) we define the problem of transaction generation and mining association rules from SW data, and (2) we propose a method to efficiently extract items and transactions suited for traditional association rules mining algorithms. This work extends that presented in [28] in the following aspects: a more elaborate and complete formalism is presented, we have updated the related work, and an exhaustive evaluation over a real scenario has been carried out.

The rest of the paper is organized as follows. Section 2 gives an overview of the related work. Section 3 explains the basics of the two integrated technologies, association rules mining and OWL DL ontologies and motivates the problem with a running example. Section 4 contains the general methodology and foundations of the approach. Section 5 shows the experimental evaluation and Section 6 gives some conclusions and future work.

2. Related work

Most research on DM for semantic data is based on inductive logic programing (ILP) [26], which exploits the underlying logic encoded in the data to learn new concepts. Some examples are presented in [23,17]. However, there is the inconvenient of rewriting the data sets into logic programming formalisms and most of these approaches are not able to identify some hidden concepts that statistical algorithms would.

Generalized association rules [32] were formerly proposed to consider item taxonomies for mining association rules. The main idea behind this method is to extend itemsets with all the ancestors of each item, then computing the frequent itemsets, and finally filtering out those rules containing an item and some of its ancestors. The resulting rules are expressed at different taxonomy detail levels, avoiding redundant rules present in the taxonomy. Clearly, this method overheads considerably the transactions length and therefore the mining algorithm. Additionally, the number of generated rules is much larger than disregarding the taxonomy. Several optimizations have been proposed to alleviate this overhead. [15] presents two association rules mining algorithms as the core of the Web personalization process. Both algorithms are efficient in the sense that they avoid the costly generation of candidate sets and the over-generalization of rules. A pattern-fragment growth method is adopted along with efficient pruning. Moreover, the special features of Web 2.0 applications are also addressed, where the taxonomies are not predefined but usually described by user-defined tags. In [36] the problem of efficiently updating the discovered generalized association rules when the item taxonomies are modified is addressed. Nevertheless, our work is not concerned with this kind of association rules but with the definition of transactions for SW data. Once the transactions are generated we can mine any kind of association rules, like generalized and quantitative ones.

Other studies extend statistical machine learning algorithms to be able to directly deal with ontologies and their associated instance data. In [8], it is shown how kernel-based machinery can encapsulate the ontology knowledge into the vector-based representation suited for support vector machines. For clustering purposes, a series of similarity (dissimilarity) measures are proposed to mine either semi-structured data [11,14] or ontological instances [12]. In the latter case, knowledge derived from the ontology is used to define the weights of the similarity measures. However, as previously mentioned, these representations are not suited for association mining, which requires to define both the set of items and the transactions from the semantic data.

Another related topic to this work is that of mining tree and graph structured data. In this line, we can find frequent subtree [10] and graph mining [20], whose aim is to identify frequent substructures in complex data sets. Albeit interesting, these algorithms do not serve the purpose of finding interesting content associations in RDF/(S) and OWL graphs because they are concerned with frequent syntactic substructures but not frequent semantically related contents. Indeed, frequent graph substructures usually hide interesting associations that involve contents represented with different detail levels of the ontology. Moreover, although the underlying structure of RDFS and OWL is a graph, reasoning capabilities must be applied to handle implicit knowledge.

The way we define transactions is similar to those proposed for analysing highly heterogeneous XML data sets. More specifically, in [22,38] it is shown that XQuery is not the most suitable query language for data extraction from heterogeneous XML data sources, since the user must be aware of the structure of the underlying documents. The lowest common ancestor (LCA) semantics can be applied to extract meaningful related data in a more flexible way. [22,38] apply some restrictions over the LCA semantics. In particular they propose SLCA [38] and MLCA [22] whose general intuition is that the LCA must be minimal. However, in [27,30] they showed that these approaches still produced undesired combinations between data items in some cases (e.g. when a data item needs to be combined with a data item at a lower level of the document hierarchy). In order to alleviate the previous limitations they propose the SPC (smallest possible context) data strategy, which relies on the notion of closeness of data item occurrences in an XML document. A similar strategy is presented in [34].

Finally, we find some work aimed at integrating knowledge discovery capabilities into SPARQL[3] by extending its grammar. Some examples are [18], which can be plugged with several DM algorithms and [19], which finds complex path relations between resources. Inspired by these works, we have also extended SPARQL grammar to define association rule patterns over the ontological data but in a less restrictive way than the one imposed by SPARQL. These patterns allow the system to focus only on the interesting features, reducing both the number and length of generated transactions.

3. Preliminaries

This section gives some background about the two integrated research areas, the semantic web and association rules, and introduces our mining problem statement in this context.

3.1. Semantic web data

Semantic web technologies are aimed at providing the necessary representation languages and tools to bring semantics to the current web contents. As a result, the W3C consortium has proposed several representation formats, all relying on XML. The resource description language (RDF) was the first language proposed by the W3C to describe semantic meta-data. In RDF there

[3] SPARQL: http://www.w3.org/TR/rdf-sparql-query,'08.