Context and Target Configurations for Mining RDF Data

Ziawasch Abedjan
Hasso-Plattner-Institut
Potsdam, Germany
ziawasch.abedjan@hpi.uni-potsdam.de

Felix Naumann
Hasso-Plattner-Institut
Potsdam, Germany
naumann@hpi.uni-potsdam.de

Categories and Subject Descriptors
H.2.8 [Database Management]: Database applications
General Terms
Algorithms

1. MINING LINKED DATA

Association rule mining has been widely studied in the context of basket analysis and sale recommendations [1]. In fact, the concept can be applied to any domain with many items or events in which interesting relationships can be inferred from co-occurrence of those items or events in existing subsets (transactions). The increasing amount of Linked Open Data (LOD) in the World Wide Web raises new opportunities and challenges for the data mining community [5]. LOD is often represented in the Resource Description Framework (RDF) data model. In RDF, data is represented by a triple structure consisting of subject, predicate, and object (SPO). Each triple represents a statement/fact.

The URI representation of subjects, predicates, and objects and their connections within statements harbor many hidden relations that might lead to new insights about the data and its domain. Resources are connected by multiple predicates, co-occurring in multiple implicit relations. At this point, frequencies and co-occurrences of statement elements become an interesting object of investigation. To mine such SPO data, several questions must be answered: What should be mined in which context, and what are the application fields for each approach. Previous work concentrates on Inductive Logic Programming and graph mining, or is restricted to scenarios where domain knowledge and complete ontology structures are available.

We propose an approach that applies association rule mining at statement level by introducing the concept of mining configurations. A configuration specifies the context of rule mining (the transactions) and the target of rule mining (the items). For each configuration, we describe the corresponding opportunities, capabilities, and application fields. As a proof-of-concept we applied our mining concept to the popular DBpedia data set [2].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SMER’11, October 28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0957-8/11/10 ...$10.00.

2. MINING RDF ON STATEMENT LEVEL

Our mining approach is based on the subject-predicate-object (SPO) view of RDF data. Any part of the RDF statement can be regarded as the context, which is used for grouping one of the two remaining parts of the statement as the target for mining. Thus, a transaction is a set of target elements associated with one context element. We call each of those combinations a configuration, and each configuration corresponds to one or more discovery use cases. Tab. 1 shows an overview of the six possible configurations. For each configuration, we describe the corresponding opportunities, capabilities, and application fields. As a proof-of-concept we applied our mining concept to the popular DBpedia data set [2].

Related Work. There is much work on mining the Semantic Web in the fields of Inductive Logic Programming and approaches that make use of the description logic of a knowledge base [6]. Those approaches concentrate on mining answer-sets of queries towards a knowledge base. Based on a general reference concept, additional logical relations are considered for refining the entries in an answer-set. This approach depends on a clean ontological knowledge base, which is usually not available. Furthermore, that approach ignores mining of rules among predicates.

As RDF data spans a graph of resources connected by predicates as edges, another related field of research is mining frequent subgraphs or subtrees [4]. However, in LOD no two different nodes in an RDF graph have the same URI unless we consider the corresponding concept of each URI. Thus, any graph mining would be restricted to concept mining and not data mining. A statistical approach for mining the Semantic Web is proposed by Nebot et al. [7], where a SPARQL endpoint allows the user to define targets of mining in any desired context. For example, their approach is able to generate rules between drugs and diseases by examining their co-occurrence in the context of different patients. Such a specific use case requires domain-specific knowledge of the existing RDF graph, which is probably not always available. Furthermore, the approach ignores the role of predicates. Among profiling tools, ProLOD is a tool for profiling LOD, which includes association rule mining on predicates for the purpose of schema analysis [3].
Table 1: Six configurations of context and target

<table>
<thead>
<tr>
<th>Context</th>
<th>Target</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Predicate</td>
<td>Schema discovery</td>
</tr>
<tr>
<td>Subject</td>
<td>Object</td>
<td>Basket analysis</td>
</tr>
<tr>
<td>Predicate</td>
<td>Subject</td>
<td>Clustering</td>
</tr>
<tr>
<td>Predicate</td>
<td>Object</td>
<td>Range discovery</td>
</tr>
<tr>
<td>Object</td>
<td>Subject</td>
<td>Topical clustering</td>
</tr>
<tr>
<td>Object</td>
<td>Predicate</td>
<td>Schema matching</td>
</tr>
</tbody>
</table>

Mining subjects in the context of objects (Conf. 5), i.e., discovering subjects that share a minimum number of object values, results in rules between entities that are topically related. Objects are values that might be associated with subjects in different relations. E.g., several persons may share the object Berlin in different roles like birth or death place or home town. In fact, up to 47 distinct predicates involve Berlin as object value. Therefore, organizations as well as persons and other concepts might share the same objects, and are consequently topically related.

In the context of objects, we face the problem of non-frequent subjects, as each subject occurs at most as often as there are predicates defined for the certain subject value. The same applies to the number of involved objects with the difference that the number of distinct objects is about 100 times higher than the number of predicates. This circumstance causes much lower support thresholds when mining in the context of objects. Again lowering the support value results at some point in generating masses of irrelevant rules. A solution might be to generalize objects or subjects values by means of their concepts or types.

Mining Predicates. While subjects represent entities in RDF data, predicates represent the schema for those entities. So, mining predicates in the context of subjects (Conf. 1) results in patterns and rules that show dependencies of schema elements among entities and can be used for schema discovery and analysis. Rules such as \{associatedBand, instrument\} → \{associatedMusicalArtist\} with 99% confidence and 3% support or \{activeYearsEndDate, party\} → \{activeYearsStartDate, birthDate, successor\} with 2.5% support and 68% confidence show that the data contains different schemata for musicians and politicians. The difference in confidence might trigger a closer examination of possible reasons for loose or tight consistency of the schemata.

Mining predicates in the context of objects (Conf. 6) aims at discovering predicates that have a strong overlap in their value ranges. As predicates define the schema of entities, rules within this configuration can be used for schema matching. For instance, we discovered rules between the predicates associatedBand and associatedMusicalArtist that have a confidence of 100% in both directions. They are caused by the fact that both predicates have exactly the same value range. On the other hand, rules such as residence → birthPlace with only 67% confidence show that even categorically similar properties might differ in their range.

Mining Objects. To pursue our view of entities and schemata, objects would represent the actual values that describe an entity within a certain schema. Thus, mining in the context of subjects (Conf. 2) means to discover patterns between values that are associated to each other by co-occurring for many entities. For example, the rule Buenos Aires → Argentina with 85% confidence shows that entities associated with a capital town are probably also associated with the corresponding country. Other rules such as Film → Television with 72% confidence show quite reasonable but not surprising associations. On the other hand, rules that correspond to entities representing ice-hockey players such as Defense → Canada with 61% confidence show also interesting dependencies regarding the player positions and their origins. Forasmuch, this configuration might represent the traditional discovery of new rules for market basket analysis.

Rules in the context of predicates (Conf. 4) imply range discovery of predicates as they connect values such as numbers, countries, or cities. Exemplary rules include \{2\} → \{2, 3\} or Albania → Italy. In fact, the mining results of this configuration is very similar to the configuration for mining subjects in the context of predicates. Regarding the fact that subjects and objects have semantically different roles in a statement, it is worth reasoning about the actual difference of both configurations.

3. CONCLUSIONS AND FUTURE WORK

We proposed a statement-level mining methodology for RDF data that is based on six basic configurations with the aid of different real examples. On this basis, further research directions include reasoning and formalizing constraints and refinements that allow unlimited configuration scenarios for different purposes. With regard to the heterogeneity of RDF, another concern of the examination is to define reasonable scopes for the algorithm parameters support and confidence. As a proof-of-concept application we are currently developing a rule network that allows auto-completion of RDF data by combining configurations for mining predicates as well as objects.

4. REFERENCES


