RDF Knowledge Graph Mining over Data Cubes

Petr Novák and Vojtěch Svátek

Prague University of Economics and Business, Czech Republic {novp19,svatek}@vse.cz

Abstract. Existing efforts in RDF knowledge graph (KG) mining solely addressed fact-level KGs. Many KGs, especially those published by government entities, however contain aggregated data modeled using data cubes. We applied a state-of-the-art KG mining tool on RDF datasets of two government institutions and on interlinked fact-level KGs (Wikidata and YAGO), in multiple settings. Some of the results are meaningful, the main challenges being the low level of support of individual rules and their difficult readability.

Keywords: knowledge graph \cdot relational mining \cdot Data Cube \cdot linked government data

1 Introduction

Algorithms for data mining from knowledge graphs (KGs) have recently been developed, which take into account the RDF data model and the effects of the open world assumption. Their prime representative is AMIE [2]. The experiments undertaken so far focused on the large datasets containing *facts* on individual entities, such as DBpedia, Wikidata or YAGO. On the semantic web we however also see *multi-dimensional* datasets capturing *aggregated* data: sums, averages, counts, etc. The RDF representation of such data, mainly published by gov-ernment institutions, relies on the Data Cube vocabulary¹ (DCV). The data follows a star pattern with a resource denoted as *observation* in the center and the 'beams' corresponding to the assignment of values to *dimensions* and *measures*. All observations from the same dataset share the same set of dimensions and measures is the observation that women in the South-Bohemian region in 2020 received an average (old-age) pension of 5.7Keuro. The dimensions here are the sex, the region and the year; the measure is the average pension.

While some effort has been already devoted to making statistical analysis of DCV-based data cubes easier [1], the presented study is, to our knowledge, the first attempt to apply the *rule mining* technology on data involving triples from multidimensional cubes. Our primary aim was to analyze the situations arising

¹ http://www.w3.org/TR/vocab-data-cube/

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

in such a scenario and the potential usability of the discovered rules. A secondary aim was to provide additional feedback to an RDF mining tool, RDFRules [6] as a novel reimplementation of AMIE developed by our group. Compared to the original AMIE, RDFRules provides, among other, ample pre-processing options, a pattern language for specifying the shape of rules, mining over multiple graphs, or rule post-processing (clustering and pruning).

2 RDF Mining Specifics

While rule mining sometimes lags behind other (statistical or neural) inductive learning paradigms in predictive accuracy, its advantage is the immediate interpretability of the models. However, rule mining algorithms, such as *Apriori* [4], are limited by the shape of the analyzed data in the form of a single or several related tables. Therefore, these algorithms are not applicable to the graphstructured RDF model. The application of Inductive Logical Programming to rule mining does not have this limitation, but requires negative examples. These are mostly unavailable in Linked Data. An algorithm specifically designed to mine rules from data operating under OWA and consisting of binary predicates (just as Linked Data) is AMIE [2]. AMIE (and its novel implementation RD-FRules [6]) mines rules consisting of a head as a single triple and a body as a conjunction of triples, with triples containing variables, e.g.

(?a ex:worksFor ?b) \land (?b ex:hasHeadquartersIn ?c) \Rightarrow (?a ex:livesIn ?c) The rules are filtered by interestingness measures, some of which are adapted from propositional mining (e.g., support) and some are new (e.g., head coverage). In the context of the Apriori algorithm, the support of a rule corresponds to the number of transactions (rows in the table) that conform to both the assumption and the conclusion of the rule. As classical support cannot be straightforwardly transferred to the case of rules with variables, for the AMIE algorithm the support measure is redefined as the number of triples that are correctly predicted by the rule. Head coverage, in turn, is a novel measure introduced by AMIE. It is defined as the ratio of the support of a rule and the head size: the number of triples having the same predicate as the rule head.

Both support and head coverage describe a quantitative significance of the rule in relation to the examined data. They quantify the true predictions of the rule but do not take into account the false predictions, as e.g. the confidence measure does. Generally speaking, confidence is a ratio of true predictions of a rule to the sum of true predictions and the counterexamples. While mining rules from the KG, however, one cannot simply consider the absence of a triple as a counterexample. AMIE treats this problem by defining a confidence measure based on so-called Partial Completeness Assumption (PCA), which states, that if the KG contains triple with a subject s a predicate p, the KG already contains a *complete* set of possible triples with the subject s and the predicate p. Therefore, a triple predicted by the rule missing in the KG is only considered a counterexample if its subject appears in the KG with the triple's predicate and with another object o'.

3 RDF Data Cube Mining Specifics

Let us now briefly compare the problem addressed in our research with both traditional analytics over data cubes and with rule mining from fact KGs.

The analysis of multi-dimensional cubes by statistical or OLAP techniques operates on continuous values of measures. Conversely, for rule mining the measures have to be discretized into intervals, to ensure both sufficient frequencies of values and a manageable search space. Since it entails some information loss, mining a single data cube after discretization would probably be meaningless. The strong side of RDF rule mining may however be the possibility to mine across multiple interlinked cubes, and further, to connect fact KGs to them.

The rules must respect the structure of the cubes. All rule atoms (triple patterns) referring to a data cube will have a variable representing the observation as their subject, and their predicates will correspond to dimensions and measures; additionally, the observation variable will be linked to the dataset IRI through the DCV predicate qb:dataSet. A further data cube can be linked to the first one through a common dimensional value; similarly, any triple from a fact KG can be linked to a dimensional value.

The presence of data cube atoms also impacts the computation of rules' interestingness measures. Namely, the total number of instantiations of a set of rule atoms whose predicates are dimensions, e.g.,

(?o ex:refArea ?a) \land (?o ex:refPeriod ?b)

equals to the product of the number of areas and the number of periods, since each combination of dimensions is present in the data cube. Cubes with more values for their dimensions thus artificially increase the support of rules in which they appear. However, as we verified [5], another interestingness measure of AMIE/RDFRules, *head coverage*, is neutral wrt. the data cube size.

The greatest problem in rule mining from aggregated data (identified in propositional rule mining [3], but valid for relational mining as well) is the assurance of *commensurability* of the dimensional values. For example, if some region is much bigger than others, it will tend to trivially appear in many rules. Decomposition of large cubes to multiple smaller ones (and then running separate mining tasks over them) is necessary in such cases; this however leads to lowered values of interestingness measures. The same discretization of measure values may also be only shared between observations with a commensurable context.

4 Experiments

We carried out a series of experiments involving datasets published in RDF by two government institutions: the pensions dataset by the Czech Social Security Administration, surveying the average amount of pension, the average age and the number of persons, and several datasets by the Czech Statistical Office, such as the 'jobs applicant and unemployment rate' (further, JAUR), or the dataset of average salaries. The shared dimensions of the datasets are the *sex*, the *year* and the *region*. Additionally, we extracted relevant portions of public fact KGs: Wikidata and YAGO. The triples extracted from Wikidata deal with heads of state and regional governments in particular years, their political parties, and the political alignment of these parties (left, center, far-right, etc.). Similarly, the triples extracted from YAGO relate to the regions.

The rules were required to have a measure in the head. The body pattern varied from task to task. Representative examples of analytical questions expressed through the RDFRules pattern functionality are: 1) Is there a relation between the pension expenses and the political alignment of the state government? 2) Is there a relation between the number of job applicants and the features of a region? 3) For which groups does a relation between salaries and pensions hold?

RDFRules was called through its API. The source code of the experiments is stored in a repository on github² in the form of jupyter notebooks.

In the first task, "political alignment vs. pensions", the pensions dataset was mined in combination with Wikidata. The data set was sliced into subcubes for each pension kind, and the discretization was performed separately on each. The rule pattern can be translated as: 'If in any year the current prime minister belongs to a political party that has a certain political alignment then the annual expenses of a certain pension kind fit into a certain interval'. The uniform size of the subcubes allowed to use the support as the interestingness measure; 29 rules had a support higher than 1. The following example rule has the support of 3 and the confidence value of 1, and states that all of the three years of the *center-right* government featured old-age pension expenses in the upper third for the period observed in the data set.

```
(?headRole x:appliesToRefPeriod ?refPeriod) ^
(wd:Q213 p:P6 ?headRole) ^
(?headRole ps:P6 ?person) ^
(?person p:P102 ?partyMembership) ^
(?alignment rdfs:label "centre-right") ^
(?party wdt:P1387 ?alignment) ^
(?partyMembership ps:P102 ?party) ^
(?partyMembership ps:P102 ?party) ^
(?partyMembership x:appliesToRefPeriod ?refPeriod) ^
(?observation cssz-dimension:refPeriod ?refPeriod) ^
(?observation cssz-dimension:druh-duchodu
cssz-pensionkind:PK_old_age_total_S_SI_SRN_ST_SD_SR_2010) ^
(?observation cssz-measure:vydaje-na-duchody-v-cr) ->
(?observation cssz-measure:vydaje-na-duchody-v-tis-kc
```

<<3.1797095155E8__3.8222294828E8)_ef3_3/3>)

In the second task, the JAUR dataset was mined in combination with YAGO. The rule pattern was more open here: any features from YAGO could appear in the body as long as they were linked to a region. The algorithm then resorted to finding common features of areas with a certain measure in a certain interval.

In the third task the two data cubes from different providers were analyzed, and their measures were required to appear one in the body and one in the head.

Across the tasks, the intuitive plausibility of the rules varied. The rules with a 'monetary' measure in the head (as in the first task) may suffer from a bias introduced by the inflation; it makes the pensions partially grow regardless the

² https://github.com/nvkp/diploma-thesis-code

government alignment. When we allowed for measures from the same cube to appear in the body and in the head, plausible but trivial rules were found (in the second task), e.g., describing the relationship between the two JAUR measures: the unemployment rate and the number of job applicants. Some of the associations found via fact KG links also featured chains reflecting rather indirect if not dubious relationships; for example, the birth regions of players of a particular sports club were associated with a high number of job applicants. The rules for the third, cross-cube task were often quite meaningful, relating, e.g., lower pensions of some kind (e.g., widow/er ones) to lower salaries, although, similar findings would presumably be discovered using statistical (e.g., regression) techniques over non-discretized data, i.e., the added value of relational mining is a bit questionable here (since no fact KG was linked in this case).

Setting the rule patterns to connect the aggregated values to the instance data of the KGs just as in the first and the second mining task, however, undoubtedly has a potential to yield new insights to the cubes' data, that cannot be obtained from performing the statistical techniques alone. And the RDFRules framework proved to be well suited for this type of analysis, thanks to its native recognition of the *owl* : *sameAs* predicates, discretization functionality, and the ability to control the shape of the generated rules with its rule pattern syntax.

5 Future Work

Evaluation of the rule quality by domain experts would be desirable in the next phase. Some smart visualization of the rules should also be considered, since the number of atoms required to express the cube structures is often high and some are not completely intuitive. Beyond the government data, the approach could also be tried on industrial data warehouses, which contain similarly structured multi-dimensional data, and their integration with both private and public KGs and subsequent discovery of cross-graph patterns might be attractive.

This work has been partially supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003).

References

- Capadisli, S., Auer, S., Riedl, R.: Linked Statistical Data Analysis. In: Semantic Statistics 2013, http://ceur-ws.org/Vol-1549/.
- Galárraga, L. A., Teflioudi, C., Hose, K., Suchanek, F.: AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In: WWW 2013.
- Chudán, D.: Association rule mining as a support for OLAP. PhD Thesis. Univ. of Econ., Prague, 2015. https://insis.vse.cz/zp/portal_zp.pl?podrobnosti_zp=25910
- Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD'93: 207-216.
- 5. Novák, P.: Experiment with rule mining from linked government data. MSc. Thesis. Prague University of Economics and Business, 2021.
- Zeman, V., Kliegr, T., Svátek, V.: RDFRules: Making RDF rule mining easier and even more efficient. *Semantic web*, 2021, Vol. 12, No. 4.