Ontologies in Support of Data Mining

Viktor Nekvapil and Ondřej Zamazal

University of Economics, W. Churchill Sq.4, 130 67 Prague 3, Czech Republic
viktor.nekvapil@vse.cz ondrej.zamazal@vse.cz

1 Introduction and Motivation

Ontologies and vocabularies are commonly used within Linked data cloud for typing individual objects. Due to their semantics, domain ontologies are used as a source of knowledge in diverse applications. Data mining phases are commonly enhanced by various forms of *background knowledge*, besides other things, to explain discovered patterns. Usually, domain expert is a source of such explanations. The aim of the proposed approach is to replace domain expert and offer explanations to the results (semi)automatically using external knowledge base and ontologies.

Let us use a discovered rule $Region(Zlinsky) \rightarrow Loan(100 - 150\rangle$[1] to illustrate our approach. The system generates explanations[2] $Region(Zlinsky) : Income(very\ low)$[3] and $Region(Zlinsky) : Infarction(very\ high)$. To compare relevancy of the explanations to the rule, we have to compare a similarity of domains of interest of terms "*Loan*" and "*Income*" to a similarity of domains of interests of terms "*Loan*" and "*Infarction*". We propose an approach to compare similarities using knowledge captured in ontologies from diverse ontology collections such as BioPortal[4] or Linked Open Vocabularies (LOV) portal.[5]

2 Offering Relevant Explanations

Offering explanations to the rules using ontologies was studied e.g. in [1]. However, in our approach, the ontologies used are not created manually, but existing ones are exploited. Our approach is described in [2]. We summarize it below and introduce a newly proposed *Ontology Based Relevancy Criterion* (OBRC).

1. We have a knowledge base with processed external data tables (called *items of knowledge*) from various domains. The *measure* in the item of knowledge is stored as per dimension[6] which we call a *connecting attribute*. Example item of knowledge is depicted in Table 1. The measure is average income in the

[1] In Region Zlinsky, amount of loan taken by a client is 100-150 thous. CZK.
[2] The approach of generating explanations is summarized in Section 2.
[3] In Region Zlinsky, the average income is very low.
[4] http://bioportal.bioontology.org/
[5] http://lov.okfn.org/dataset/lov/
[6] The terms *measure* and *dimension* are used as in the business intelligence domain.

Czech republic in 2016 (shortly 'Income'), stored per dimension (*connecting attribute*) Region.[7]

2. We have a set of association rules output by a data mining system. Each rule contains a *connecting attribute* (*Region* in the example) and other attributes.

Table 1. Item of knowledge – Average income in thousands CZK/month per Region

Region (connecting attribute)	Income (measure)	Rank	Level
Prague	35,1	1	Very high
Stredocesky	27,3	2	Very high
Plzensky	26,0	3	High
...	...	...	...
Zlinsky	23,8	13	Very low
Karlovarsky	22,7	14	Very low

3. A row of the item of knowledge is retrieved as an *explanation* (in a form, e.g., $Region(Zlinsky) : Income(very\ low)$) to the rule if:
   (a) The connecting attribute value in the rule matches the connecting attribute value of the row ($Region(Zlinsky)$ in our example).
   (b) The row has a very high or very low level of measure.[8]
   (c) The row satisfies the *Ontology Based Relevancy Criterion* (OBRC).

The motivation of enhancing the current method [2] (adding the OBRC – point *3(c)* above) is the fact that the knowledge base includes thousands of items of knowledge and the need is to asses their relevancy to the output rule. The proposed approach of the OBRC relies on the following assumptions:

1. Ontologies express knowledge about a certain domain of interest.
2. It is possible to decide whether the terms from the rule and explanation are present in a certain ontology.
3. If both terms from the rule and explanation are present in one ontology, they are semantically connected, i.e., they are possibly from the same domain and thus the explanation is relevant.[9]
4. The more common ontologies, the more relevant the explanation is.

The heuristic described above can be implemented using existing tools, e.g., the LOV portal and its terms search service.[10] LOV is used as one of the sources

[7] The items of knowledge are extracted from the publicly available sources (statistical offices, ministries etc.). Nonetheless, it is possible to use any external data.

[8] Here, this simulates the usual way of thinking of domain experts where the very high or very low levels are interesting. The levels of the rows can be assigned either by domain expert or using automated means based on ranking of the rows or based on more sophisticated statistical methods.

[9] There can be more than one attribute in the rule; we compute common ontologies for each pair *term from rule–term from explanation* and weight it by number of pairs.

[10] `http://lov.okfn.org/dataset/lov/terms`

of the OOSP tool (Online Ontology Set Picker) [3] where one can use ontology search based on lexical tokens. OOSP has been selected[11] since it allows to search for ontologies from various ontology collections such as BioPortal, LOV and NanJing.[12] In order to support this task we prepared a REST-based service for searching of ontologies within OOSP. There are five parameters: *word* to be searched in ontologies, *collection* specifying a set of ontologies for searching in, *comparator* specifying exact or token (fulltext) based search, *entities* stating which kind of entities (classes, properties etc.) should be considered, *scope* means which part of ontology should be searched through such as local names, labels or comments.[13]

3 The Usage Example

Let us demonstrate the approach on the real data set from the *financial services industry (FSI)* domain, items of knowledge being from *FSI* (Flat price, Income, Mortgage) and *healthcare* domain (Diabetics,[14] Abortions[15]). The data mining process generated several rules, one of them being $Region(Vysocina) \rightarrow Loan(100 - 150\rangle$,[16] three candidate explanations[17] are retrieved:

1. $Region(Vysocina) : Mortgage(very\ low)$[18]
2. $Region(Vysocina) : Diabetics(very\ low)$
3. $Region(Vysocina) : Abortions(very\ low)$

As obvious, the first explanation is relevant to the resulting rule, i.e., it is from the FSI domain so as the attribute Loan from the resulting rule. Thus, it could be useful for the end-user as an additional knowledge helping to explain the resulting rule. The explanations (2) and (3) are from completely different domain (healthcare) and this should be reflected by the OBRC.

The results of the requests on the *OOSP service* are presented in Table 2. As ontology pool we use a set of ontologies from LOV (Nov. 2017 snapshot), OntoFarm, NanJing and Bioportal (Nov. 2017 snapshot) available from the OOSP service.[19]

All the 3 requests yield 2 common ontologies for the terms. To take also the frequency of the terms from explanations (term 2) into account, relative frequency (*% of common ontologies – %CO*) is computed. For request #1, from the 3 ontologies containing term “mortgage”, 2 of them are common with the

[11] However, any other similar tool for an ontology search could be used.

[12] `http://ws.nju.edu.cn/njvr/`

[13] For example, `https://owl.vse.cz/OOSPservices/api/v1/search3?word=vehicle&collection=all1&comparator=token&entities=7&scope=15`

[14] Average number of diabetics per regions per 100.000 inhabitants.

[15] Number of abortions per regions per 100 newborns.

[16] In Region Vysocina, amount of loan taken by a client is 100-150 thous. CZK.

[17] They satisfy both the conditions 3(a) and 3(b) stated above.

[18] In Region Vysocina, the average amount of mortgage taken is very low.

[19] We also included Bioportal because of the healthcare domain.

Table 2. Results of the requests on OOSP API

#	Term 1 (from rule)	Term 2 (from explanation)	No. of ontologies for term 1	No. of ontologies for term 2	No. of common ontologies	% of common ontologies (%CO)	
1	Loan	Mortgage	13	3	2	2/3 = 0.66	**66 %**
2	Loan	Diabetics	13	28	2	2/28 = 0.07	**7 %**
3	Loan	Abortions	13	34	2	2/34 = 0.06	**6 %**

term "loan" (which is 66%). Intuitively, this seems as a very strong correlation between the two terms. For term 'diabetics' and 'abortions' the $\%CO$ is much lower (7% and 6% respectively). We can then score the explanations according to $\%CO$ and return, for example, only the top 3 to the user using some minimum threshold for $\%CO$ (e.g. 50 %). Based on this example, we can say that the OBRC might have the ability to distinguish between different domains and thus improving the relevancy of the retrieved explanations. To support this fact, additional experiments need to be conducted.

4 Discussions, Conclusions and Future Work

It is assumed that the knowledge base will be shared across the users from different domains. This approach has several limitations. Very general ontologies can include practically any concept even from different domains – this limitation can be mitigated by the fact that the majority of ontologies in the collection are domain-specific. Further, the names of the terms have to be related to its semantic meaning, which could be problematic in case of encoded column names (e.g., CF_108). Next, there can be natural language problems: e.g., different languages, different forms of one word (singular, plural etc.), homonyms etc.

To conclude, the OBRC is useful in situations where only one word is used for naming both the attribute and the measure. This word needs to have a semantic meaning to utilize the OBRC. To enlarge the ability to find relevant explanations, more sophisticated algorithms dealing with linguistic problems need to be deployed. That is true for the multi-word names and morphological forms of words. For increasing the probability of finding relevant terms, dictionary of synonyms (e.g., WordNet)[20] could be used. The implementation of those algorithms is left as a future work. More advanced methods of measuring the relevancy (e.g., measuring distance between the two terms) will be developed in future.

Acknowledgement OZ has been partly supported by the CSF grant no. 18-23964S.

References

1. Svátek V., Rauch J., Flek M.: Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality. In: ECML/PKDD05. 2005.
2. Nekvapil, V.: Data Mining with Trusted Knowledge. In: FedCSIS. 2017.
3. Zamazal O., Svátek V.: OOSP: Ontological Benchmarks Made on the Fly. In: SumPre Workshop at ESWC 2015. Portoroz, Slovenia. 2015.

[20] `https://wordnet.princeton.edu/`