User-Driven Pattern Mining on knowledge graphs: an Archaeological Case Study

Wilcke, W.X.; de Boer, Viktor; van Harmelen, Frank

2017

document version
Publisher's PDF, also known as Version of record

User-Driven Pattern Mining on knowledge graphs: an Archaeological Case Study

Wilcke, WX W.X.WILCKE@VU.NL
Department of Computer Science,
Department of Spatial Economics,
VU University Amsterdam, The Netherlands

de Boer, V V.DE.BOER@VU.NL
van Harmelen, FAH FRANK.VAN.HARMELEN@VU.NL
Department of Computer Science,
VU University Amsterdam, The Netherlands

Keywords: Knowledge Graph, Pattern Mining, Hybrid Evaluation, Digital Humanities, Archaeology

Abstract

In this work, we investigate to what extent data mining can contribute to the understanding of archaeological knowledge, published as knowledge graph, and which form would best meet the communities' needs. A case study was held which involved the user-driven mining of generalized association rules. Experiments have shown that the approach yielded mostly plausible patterns, some of which were rated as highly relevant by domain experts.

1. Introduction

Digital Humanities communities have recently began to show a growing interest in the *knowledge graph* as data modelling paradigm (Hallo et al., 2016). In this paradigm, knowledge is encoded as edges between vertices and is supported by semantic background knowledge. Already, many humanity data sets have been published as such, with large contributors being European archaeological projects such as CARARE and ARIADNE. These data have been made available in the *Linked Open Data* (LOD) cloud – an internationally distributed knowledge graph – bringing large amounts of structured data within arm's reach of archaeological researchers. This presents new opportunities for data mining (Rapti et al., 2015).

Appearing in *Proceedings of Benelearn 2017*.

In this work[1], we have investigated to what extent data mining can contribute to the understanding of archaeological knowledge, published as knowledge graph, and which form would best meet the communities' needs. For this purpose, we have constructed a pipeline which implements a state-of-the-art method to mine generalized association rules directly from the LOD cloud in an overall user-driven process (Freitas, 1999). Produced rules take the form: $\forall\chi(Type(\chi,t) \rightarrow (P(\chi,\phi) \rightarrow Q(\chi,\psi)))$. Their interestingness has been evaluated by a group of raters.

2. Approach

Our pipeline[2] facilitates the rule mining algorithm, various pre- and post-processing steps, and a simple rule browser. We will briefly touch on the most crucial components next:

Data Retrieval: On start, users are asked to provide a target pattern which defines their specific interest, e.g., ceramic artefacts. Optionally, users may specify numerous parameters which, if left empty, are set to defaults. Together, these are translated into a query which is used to construct an in-memory graph from the data retrieved from the LOD cloud.

Context Sampling: Entities that match the supplied target pattern (i.e., target entities) are ex-

[1]This research has been partially funded by the ARIADNE project through the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193.

[2]Available at `github.com/wxwilcke/MINOS`.

tended with other entities related to them: their context. Unless specified by the user, contexts are sampled breath-first up to a depth of 3. This results in n subgraphs, with n equal to the total number of target entities in the in-memory graph. These subgraphs can be thought of as analogous to the instances in tabular data sets.

Pattern Mining: Our pipeline implements SWARM: a state-of-the-art generalized association rule mining algorithm (Barati et al., 2016). We motivate its selection by the algorithm's ability to exploit semantic background knowledge to generalize rules. In addition, the algorithm is transparent and yields interpretable results, thus fitting the domain requirements (Selhofer & Geser, 2014).

Dimension Reduction: A data-driven evaluation process is used to rate rules on their commonness. Hereto, we have extended the basic support and confidence measures with those tailored to graphs. Rules which are too rare or too common rules are omitted from the final result, as well as those with omnipresent relations (e.g., type and label). Remaining rules are shown in a simple faceted rule browser, which allows users to interactively customize templates (Klemettinen et al., 1994). For instance, to set acceptable ranges for confidence and support scores, as well as to specify the types of entities allowed in either or both antecedent and consequent.

3. Experiments

Experiments were run on an archaeological subset (±425k facts) of the LOD cloud[3], which contains detailed summaries about archaeological excavation projects in the Netherlands. Each summary holds information on 1) the project's organisational structure, 2) people and companies involved, 3) reports made and media created, 4) artefacts discovered together with their context and their (geospatial and stratigraphic) relation, and 5) fine-grained information about various locations and geometries.

Four distinct experiments have been conducted, each one having focussed on a different granularity of the data: A) project level, B) artefact level, C) context level, and D) subcontextual level. These were chosen together with domain experts, who were asked to describe the aspects of the data most interesting to them.

Results and Evaluation

Each experiment yielded more than 35,000 candidate rules. This has been brought down to several thousands using the aforementioned data-drive evaluation process. The remaining rules were then ordered on confidence (first) and support (second).

[3]Available at `pakbon-ld.spider.d2s.labs.vu.nl`.

Table 1. Normalized separate and averaged plausibility values (nominal scale) for experiments A through D as provided by three raters ($\kappa = -1.28e^{-3}$).

Experiment \ Rater	1	2	3	Mean
A	1.00	1.00	0.00	**0.67**
B	0.80	0.80	0.00	**0.53**
C	0.80	0.80	0.20	**0.60**
D	1.00	1.00	0.80	**0.93**
Mean	**0.90**	**0.90**	**0.25**	0.68

Table 2. Normalized separate and averaged relevancy values (ordinal scale) for experiments A through D as provided by three raters ($\kappa = 0.31$).

Experiment \ Rater	1	2	3	Mean
A	0.13±0.18	0.13±0.18	0.00±0.00	**0.09**±0.12
B	0.53±0.30	0.53±0.30	0.33±0.47	**0.47**±0.36
C	0.53±0.30	0.33±0.24	0.67±0.41	**0.51**±0.32
D	0.60±0.28	0.47±0.18	0.80±0.45	**0.62**±0.30
Mean	**0.45**±0.31	**0.37**±0.26	**0.45**±0.48	0.42±0.35

For each experiment, we selected 10 example rules from the top-50 candidates to create an evaluation set of 40 rules in total. Three domain experts were then asked to evaluate these on both plausibility and relevancy to the archaeological domain. Each rule was accompanied by a transcription in natural language to further improve its interpretability. For instance, a typical rule might state: *"For every artefact in the data set holds: if it consists of raw earthenware (Nimeguen), then it dates from early Roman to late Roman times"*.

The awarded plausibility scores (Table 1) indicate that roughly two-thirds of the rules (`0.68`) were rated plausible, with experiment D yielding the most by far. Rater 3 was far less positive than rater 1 and 2, and has a strong negative influence on the overall plausibility scores. In contrast, the relevancy scores (Table 2) are in fair agreement with an overall score of `0.42`, implying a slight irrelevancy. This can largely be attributed to experiment A, which scored considerably lower than the other experiments.

4. Conclusion

Our raters were positively surprised by the range of patterns that we were able to discover. Most of these were rated plausible, and some even as highly relevant. Nevertheless, trivialities and tautologies were also frequently encountered. Future research should focus on this by improving the data-driven evaluation step.

References

Barati, M., Bai, Q., & Liu, Q. (2016). *Swarm: An approach for mining semantic association rules from semantic web data*, 30–43. Cham: Springer International Publishing.

Freitas, A. A. (1999). On rule interestingness measures. *Knowledge-Based Systems*, *12*, 309–315.

Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2016). Current state of linked data in digital libraries. *Journal of Information Science*, *42*, 117–127.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. I. (1994). Finding interesting rules from large sets of discovered association rules. *Proceedings of the third international conference on Information and knowledge management* (pp. 401–407).

Rapti, A., Tsolis, D., Sioutas, S., & Tsakalidis, A. (2015). A survey: Mining linked cultural heritage data. *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)* (p. 24).

Selhofer, H., & Geser, G. (2014). *D2.1: First report on users needs* (Technical Report). ARIADNE. `http://ariadne-infrastructure.eu/Resources/D2.1-First-report-on-users-needs`.