Improving RDF data through semantic association rules mining

Addi Ait-Mlouk, Xuan-Son Vu, Erik Elmroth, Lili Jiang

[1]Department of Computing Science, Umeå University, Sweden;
[1]{addia, sonvx, elmroth, lili.jiang}@cs.umu.se

Abstract. The ultra-connected world has been generating massive volumes of heterogeneous data stored in different data sources. And these data sources need to be normalized and interconnected to create a linked open data (LOD) that can be used to analyze, extract useful semantic knowledge, and present it as a valid element for decision making. Semantic web techniques (e.g., RDF, SPARQL) have been widely used for knowledge discovery. However, the primary issue of the semantic web is insufficiently integrated approaches, incompleteness, and incorrectness. To tackle this issue, we applied Natural Language Processing (NLP) and data mining to propose an approach for extracting semantic association rules from text stored in RDF data. The proposed approach is applied to mypersonality data and allows users to process status update, extract entities (NER), and then generate semantics transactions for traditional association rules algorithms.

Keywords: RDF, NLP, data mining, semantic association rules

1 Introduction

Linked Open Data (LOD) is mostly presented in the RDF triple (SPO). However, these data are incomplete, it requires a promising technique to explore and extract new facts from text format. Association rule mining is a promising approach to generate such data, as we show in this article. We propose an approach that applies association rule mining at RDF data (status update from mypersonality *) by using NLP techniques and data mining to generate new facts from text data that can be used to enrich and improve mypersonnality knowledge base.

NLP is a research field of artificial intelligence, and computational linguistics, It aims to ensure interactions between computers and human natural languages. It concerned with data represented as unstructured text to explores how computers can be used to understand and manipulate natural language text to perform desired tasks. NLP has been applied in a number of fields of study, such as machine translation, speech recognition, text mining, multilingual information retrieval, and artificial intelligence.

*https://www.psychometrics.cam.ac.uk/productsservices/mypersonality

2 Related Work

Recently, many researchers combine the Semantic Web (SW) and data mining techniques to improve RDF data and knowledge graph. Most related researches on mining SW are focused on Inductive Logic Programming (ILP) such as WARMER [6] and ALEPH [7], these approaches are based on ILP to generate association rules. Galarraga et al.[5, 4] proposed an approach called AMIE and AMIE+ to generate closed association rules from RDF. Moreover, Molood et al. proposed a new approach called SWARM [3], this approach is based on AMIE introduce anthologies and consider both the knowledge from schema level and instance level to enrich and classify extracted semantic association rules. The SW is represented as subject, predicate, and object, where subject and object are resources and literals whereas predicate present the relationship between subject and object. In this short paper, our proposed approach takes advantage of NLP techniques and data mining to improve RDF data through association rules mining. In the following section, we first introduce NLP and then describe the proposed approach.

3 Proposed approach

One important application area that is relatively new is the social network that refers to the process of exploring social structures through the use of network and graphs. NLP is employed to extract information about different types of entities, relationships or events by following approaches such as Summarization, Chunking, Part-of-speech tagging, Named Entity Recognition, Named Entity Disambiguation, Relation Extraction, and Sentiment Analysis. The text is processed by using automatic Summarization that is the process of cleansing a text document in order to create a summary that keeps the most significant data. Furthermore, the text is analyzed by applying tokenization that is a process of splitting the text into a string of characters called tokens. These tokens can be analyzed by using stemming, lemmatization and POS, where, POS is a process of assigning morphological tags to each token, and the stemming is an approach that removing the suffixes of the token in order to give a good approximation to the word.

NER. The extraction of information from the social networks (Facebook status updates) is a very important task in order to construct a knowledge graph about user personality, political view, targeting for marketing services, etc. In a social network, there is a need for exploring and analyzing user data by using data mining and NLP that generate entities of interest. For this purpose, different techniques such as preprocessing and automatic memorization are used. The extraction of entities from status update (mypersonality) is a very important task for different research areas because they are related to personality, emotional and sentiment analysis. The approach proposed is based on three general strategies: preprocessing, NLP pipeline, mining configuration [1], and extraction of semantic transaction. Figure 6 present an overview of the approach.

Association rules. Was initiated by Agrawal [2] to analyze transactional databases. It usually defined as an implication of the form $A \rightarrow B$ such as

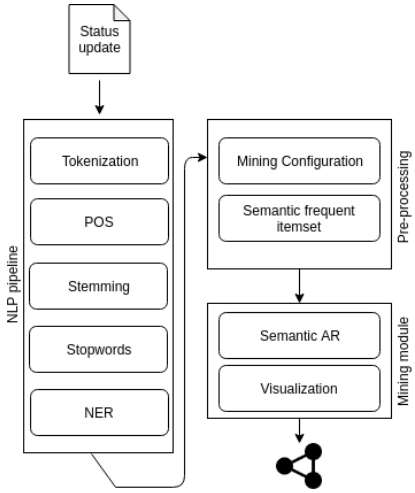

Fig. 1: Mining SAR from status update

$A, B \subset I$ and $A \cap B = \emptyset$. In order to select interesting rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence.

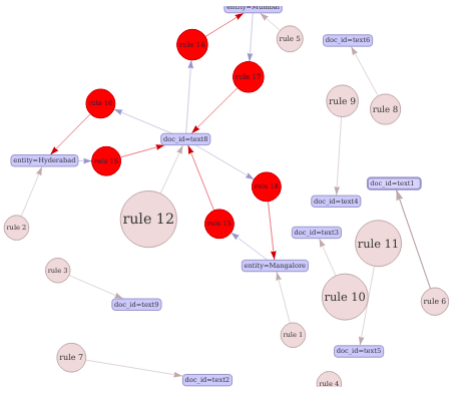

Fig. 2: Semantic association rules

By using this approach on mypersonality (status update) we extracted new semantic association rules compared to the work presented in [8]. The result shows new relationships between different Object and subject (OS), these relations can be used to improve RDF data and generate new facts for knowledge base completion and KG representation.

4 Conclusion

In this paper, we have proposed an approach to reveal semantic transaction from text data from RDF by considering NLP techniques and mining configuration. The existing approach focuses on triple (Subject, Predicate, object) to generate semantic association rules and ignore text data. To solve this issue we proposed a new approach to extract SAR from text data. this approach is automatically able to extract the configuration from text and construct semantic transaction. We applied this approach to mypersonality data(status update) to process status for different users and extract entities to be considered by the mining step to generate semantic association rules for improving KB and KG representation.

Acknowledgement

This work is supported by the Federated Database project funded by Umeå University, Sweden. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at HPC2N center. The authors also thank the myPersonality project for data contribution.

References

1. Abedjan, Z., Naumann, F.: Improving rdf data through association rule mining. Datenbank-Spektrum **13**(2), 111–120 (Jul 2013). https://doi.org/10.1007/s13222-013-0126-x, https://doi.org/10.1007/s13222-013-0126-x
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. pp. 487–499. VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994), http://dl.acm.org/citation.cfm?id=645920.672836
3. Barati, M., Bai, Q., Liu, Q.: Mining semantic association rules from rdf data. Knowledge-Based Systems **133**, 183 – 196 (2017). https://doi.org/https://doi.org/10.1016/j.knosys.2017.07.009, http://www.sciencedirect.com/science/article/pii/S0950705117303258
4. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with amie$$+$$+. The VLDB Journal **24**(6), 707–730 (Dec 2015). https://doi.org/10.1007/s00778-015-0394-1, https://doi.org/10.1007/s00778-015-0394-1
5. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: Amie: Association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the 22Nd International Conference on World Wide Web. pp. 413–422. WWW '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2488388.2488425, http://doi.acm.org/10.1145/2488388.2488425
6. Goethals, B., Van Den Bussche, J.: Relational association rules: Getting w armer. Pattern Detection and Discovery pp. 145–159 (2002), cited By 2
7. Muggleton, S.: Inverse entailment and progol. New Generation Computing **13**(3), 245–286 (Dec 1995). https://doi.org/10.1007/BF03037227, https://doi.org/10.1007/BF03037227
8. Xuan-Son Vu, Addi Ait-Mlouk, E.E.L.J.: Graph-based interactive data federation system for heterogeneous data retrieval and analytics. In: Demo Track, In: Proceedings of the The Web Conference 2019 (to appear). TheWebConf '19 - formerly WWW, International World Wide Web Conferences Steering Committee (2019)