Exploiting Semantic Web Knowledge Graphs in Data Mining

Petar Ristoski

IBM Research Almaden, CA, US
petar.ristoski@ibm.com

Abstract. Data Mining and Knowledge Discovery in Databases (KDD) is a research field concerned with deriving higher-level insights from data. The tasks performed in that field are knowledge intensive and can often benefit from using additional knowledge from various sources. Therefore, many approaches have been proposed in this area that combine Semantic Web data with the knowledge discovery process. Semantic Web knowledge graphs contain factual knowledge about real word entities and the relations between them, which can be utilized in various natural language processing, information retrieval, and any data mining applications. In this thesis, we investigate the hypothesis if Semantic Web knowledge graphs can be exploited as background knowledge in different steps of the knowledge discovery process, and different data mining tasks. More precisely, we aim to show that Semantic Web knowledge graphs can be utilized for generating valuable data mining features that can be used in various data mining tasks. In this thesis, we present several unsupervised approaches for identifying, collection, and integrating useful background knowledge from Semantic Web knowledge graphs, ranging from simple feature engineering approaches to RDF2vec embeddings. Furthermore, we showcase the use of KGs in a number of data mining applications, such as recommender sytems, entity and document modeling, and taxonomy induction.

1 Introduction

Semantic Web knowledge graphs are a backbone of many information systems that require access to structured knowledge. Those knowledge graphs contain factual knowledge about real world entities and their relations and attributes in a fully machine-readable format. Following the principles of the Semantic Web [3], such knowledge graphs are publicly available as Linked Open Data [4]. Linked Open Data (LOD) is an open, interlinked collection of datasets in machine-interpretable form, covering multiple domains from life sciences to government data [25]. Some of the most used Semantic Web knowledge bases are DBpedia [2], YAGO [26], and Wikidata [28]. In the last decade, a vast amount of approaches have been proposed that combine methods from knowledge discovery with Semantic Web knowledge graphs. The goal of those approaches is to support different data mining tasks, or to improve the Semantic Web itself.

In their seminal paper from 1996, Fayyad et al. [8] introduced a process model for knowledge discovery processes. The model comprises five steps, which lead from raw data to actionable knowledge and insights which are of immediate value to the uses: *Selection*, *Preprocessing*, *Transformation*, *Data Mining*, and *Evaluation and Interpretation*. It has been shown that Semantic Web knowledge graphs can support each step of the KDD pipeline. Given a set of local data (such as a relational database), the first step is to link the data to the corresponding knowledge graph concepts from the chosen LOD dataset. Once the local data is linked to a LOD dataset, we can explore the existing links in the dataset pointing to the related entities in other LOD datasets. In the next step, various techniques for data consolidation, preprocessing and cleaning are applied, e.g., schema matching, data fusion, value normalization, treatment of missing values and outliers. Next, some transformations on the collected data need to be performed in order to represent the data in a way that it can be processed with any arbitrary data analysis algorithms. Since most algorithms demand a propositional form of the input data, this usually includes a transformation of the graph-based LOD data to a canonical propositional form. After the data transformation is done, a suitable data mining algorithm is selected and applied on the data. In the final step, the results of the data mining process are presented to the user. Here, to ease the interpretation and evaluation of the results of the data mining process, Semantic Web and LOD can be used as well.

In this thesis, we show how Semantic Web knowledge graphs can be used in all the steps of the KDD pipeline, with the main focus being the third step of knowledge discovery process, *Transformation*, proposing multiple feature propositionalization strategies, which are evaluated on a large set of datasets, and applied in real world applications. The contribution of this thesis is broad and diverse, showing the potential of Semantic Web knowledge graphs as background knowledge in data mining. In particular, this thesis makes the following contributions:

- In-depth overview and comparison of existing approaches for exploiting Semantic Web knowledge graph in each step of the knowledge discovery pipeline, and data mining in general [20].
- A collection of benchmark datasets for systematic evaluations of machine learning tasks using background knowledge from Semantic Web knowledge graphs [23].
- Empirical evaluation of propositionalization strategies for generating features from knowledge graphs [17,13].
- An approach for feature selection in hierarchical feature spaces [18], which can be used with class hierarchies in ontologies.
- A tool for mining the web of Linked Data, providing approaches for generating, selecting and integrating features from many LOD sources.
- An approach for Semantic Web knowledge graphs embeddings, and their applications in data mining [21,6].
- A list of developed applications that exploit knowledge graphs, i.e., a tool for analyzing statistics with background knowledge from LOD [16,19]; Three

recommender system approaches [15,24,22]; An approach for entities and document modeling [21]; and an approach for taxonomy induction [14].

2 Approach

RapidMiner LOD Tool: While many domain-specific applications use Linked Open Data, general-purpose applications rarely go beyond displaying the mere data, and provide little means of deriving additional knowledge from the data. At the same time, sophisticated data mining platforms exist, which support the user to find patterns in data, providing meaningful visualizations, etc. What is missing is a *bridge* between the vast amount of data on the one hand, and intelligent data analysis tools on the other hand. In this thesis, we have developed an extension for the state of the art data mining environment RapidMiner,[1] which allows for bridging the gap between the Web of Data and data mining, and can be used for carrying out sophisticated analysis tasks on and with LOD. The extension provides means to automatically connect local data to background knowledge from LOD, or load data from the desired LOD source into the RapidMiner platform, which itself provides more than 400 operators for analyzing data, including classification, clustering, and association analysis.

Feature Generation and Selection: Most data mining algorithms work with a propositional *feature vector* representation of the data, i.e., each instance is represented as a vector of numerical or nominal features. Linked Open Data, however, comes in the form of *graphs*, connecting resources with types and relations, backed by a schema or ontology. Thus, for accessing Semantic Web knowledge graphs with existing data mining tools, transformations have to be performed, which create propositional features from the graphs, i.e., a process called *propositionalization* [9]. In this thesis, we develop five strategies for creating features from types and relations in LOD [17]. However, for many practical applications, the set of features can be very large, which leads to problems both with respect to the performance as well as the accuracy of learning algorithms. Thus, it is necessary to reduce the set of features in a preprocessing step, i.e., perform a *feature selection*. To address this problem, in this thesis, we utilize the fact that many knowledge graphs use detailed hierarchies for classifying instances, and introduce an approach that exploits those hierarchies for feature selection [18].

Semantic Web Knowledge Graph Embeddings: As the previously proposed feature generation strategies do not scale when the input dataset is large, i.e., the number of generated features quickly becomes unmanageable and requires feature selection, we introduce an approach for Semantic Web knowledge graphs embedding. In language modeling, vector space word embeddings have been proposed by Mikolov et al. [11]. They train neural networks for creating a low-dimensional, dense representation of words, which show two essential properties: (a) similar words are close in the vector space, and (b) relations between pairs of

[1] `http://www.rapidminer.com/`

words can be represented as vectors as well, allowing for arithmetic operations in the vector space. In this work, we adapt those language modeling approaches for creating a latent representation of entities in RDF graphs, called RDF2Vec [21]. Since language modeling techniques work on sentences, we first convert the graph into a set of sequences of entities using two different approaches, i.e., random graph walks and Weisfeiler-Lehman Subtree RDF graph kernels [29]. In the second step, we use those sequences to train a neural language model, which estimates the likelihood of a sequence of entities appearing in a graph. Once the training is finished, each entity in the graph is represented as a vector of latent numerical features. We show that the properties of word embeddings also hold for RDF entity embeddings, and that they can be exploited for various tasks. The generation of the entities vectors is task and dataset independent, i.e., once the vectors are generated, they can be used for machine learning tasks, like classification and regression. Furthermore, since all entities are represented in a low dimensional feature space, building the learning models and algorithms becomes more efficient. In a later work, we extend the RDF2Vec approach by introducing approaches to direct the random walks in more meaningful ways, i.e., being able to capture more important information about each entity in the graph [6]. We propose 12 weighting strategies which influence the walks and, thus, the resulting sequences. The weighing strategies are based on the predicate frequency, object frequency, predicate-object frequency and general node popularity scores, like PageRank [27]. The experiments show that the choice of weights has a crucial influence on the utility of the resulting embeddings.

3 Evaluation

In the recent years, several approaches for machine learning on the Semantic Web have been proposed. However, no extensive comparisons between those approaches have been undertaken, in particular due to a lack of publicly available benchmark datasets. As part of this thesis, we developed a collection of 22 benchmark datasets of different sizes, derived from existing LOD datasets as well as from external classification and regression problems linked to the LOD cloud. Such a collection of datasets can be used to conduct quantitative performance testing and systematic comparisons of approaches, which also allows for determining the statistical significance of the findings [23]. We use this collection of datasets to perform extensive evaluation of the above mentioned feature generation approaches, i.e., simple propositionalization approaches [17,13], and graph embedding approaches [21,6]. We compare our approaches to the state-of-the-art graph embedding approaches, i.e., TransE [5], TransH [30] and TransR [10], and graph kernel approaches, i.e., Weisfeiler-Lehman graph kernel and Intersection Tree Path kernel for RDF [29]. The feature vectors are applied in two learning tasks, i.e., classification and regression, using different learning algorithms. For comparing the approaches, we follow the approach introduced by Demšar [7], which allows for systematic comparisons including tests for statistical significance

The results show that our approach consistently outperforms the related approaches on all the datasets [21,6].

Furthermore, our approaches have been extended and applied to a variety of other applications, evaluated on the benchmark datasets in the literature for the specific task, i.e., recommender systems [15,24,22,21], entity and document modeling [21], taxonomy induction [14] and event-based knowledge reconciliation [1]. Furthermore, there is a number of third party applications that utilize the approaches described in this thesis [12].

4 Conclusions and future work

With the increased number of size of general purpose knowledge graphs, information contained therein can be used in the data mining and knowledge discovery process. However, the knowledge discovery process is still not tapping the full potential that is provided by the Semantic Web. In this thesis, we developed several approaches that exploit Semantic Web knowledge graphs in order to aid different steps of the knowledge discovery process. Extensive evaluation of the approaches showed significant improvement over related approaches, and the vast amount of applications we built on top of the approaches shows the significance of this thesis for real-world use-cases. To that end, the work presented in this thesis has been published in 6 international journals and 13 international conferences, resulting in more than 800 citations and h-index of 14. The full thesis can be found here [12].

Future Work With the recent developments, knowledge graphs are becoming a mainstream representation paradigm for both open and corporate data. This leads to a rapid development of approaches for managing such data. i.e., generating, consuming and mining. One of the future research directions is converting the vast amount of unstructured data published on the Web in various formats, e.g., text, lists, tables, etc, into graph data. This data contains useful knowledge that aligned with existing Semantic Web knowledge graphs can be of a high value. Furthermore, maintaining high quality knowledge graphs is not trivial. To address this issues, data mining solutions can be used as well. While, in this thesis we described a set of approaches for converting graph data in propositional form that can be directly used in existing data mining tools and algorithms, in future work, we believe more effort will be put in developing data mining algorithms that will be applicable directly on graph data. When developing such approaches, a lot of effort will go into optimizing the data mining algorithms, due to the large size of the knowledge graphs. In many cases multiple graphs need to be merged together, which exponentially increases the computational complexity. This brings the attention to the need for developing efficient approaches for merging large-scale knowledge graphs.

To summarize, we believe that contributions like this thesis will advance the technology for graph-based representation and consumption of data, which will integrate data mining as a non-separable part of the maintenance and consumption of the data.

References

1. Alam, M., Recupero, D.R., Mongiovi, M., Gangemi, A., Ristoski, P.: Event-based knowledge reconciliation using frame embeddings and frame similarity. Knowledge-Based Systems 135, 192–203 (2017)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)
3. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american 284(5), 28–37 (2001)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data: The story so far. In: Semantic services, interoperability and web applications: emerging concepts, pp. 205–227. IGI Global (2011)
5. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems. pp. 2787–2795 (2013)
6. Cochez, M., Ristoski, P., Ponzetto, S.P., Paulheim, H.: Biased graph walks for rdf graph embeddings. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics. p. 21. ACM (2017)
7. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research 7(Jan), 1–30 (2006)
8. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM 39(11), 27–34 (1996)
9. Kramer, S., Lavrač, N., Flach, P.: Propositionalization approaches to relational data mining. In: Relational data mining, pp. 262–291. Springer (2001)
10. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
12. Ristoski, P.: Exploiting semantic web knowledge graphs in data mining. Ph.D. thesis (2018)
13. Ristoski, P., Bizer, C., Paulheim, H.: Mining the web of linked data with rapidminer. Web Semantics: Science, Services and Agents on the World Wide Web 35, 142–151 (2015)
14. Ristoski, P., Faralli, S., Ponzetto, S.P., Paulheim, H.: Large-scale taxonomy induction using entity and word embeddings. In: Proceedings of the International Conference on Web Intelligence. pp. 81–87. ACM (2017)
15. Ristoski, P., Mencía, E.L., Paulheim, H.: A hybrid multi-strategy recommender system using linked open data. In: Semantic Web Evaluation Challenge. pp. 150–156. Springer (2014)
16. Ristoski, P., Paulheim, H.: Analyzing statistics with background knowledge from linked open data. In: Workshop on Semantic Statistics. vol. 140 (2013)
17. Ristoski, P., Paulheim, H.: A comparison of propositionalization strategies for creating features from linked open data. Linked Data for Knowledge Discovery 6 (2014)
18. Ristoski, P., Paulheim, H.: Feature selection in hierarchical feature spaces. In: International Conference on Discovery Science. pp. 288–300. Springer (2014)

19. Ristoski, P., Paulheim, H.: Visual analysis of statistical data on maps using linked open data. In: European Semantic Web Conference. pp. 138–143. Springer (2015)
20. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: A comprehensive survey. Web semantics: science, services and agents on the World Wide Web 36, 1–22 (2016)
21. Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., Paulheim, H.: Rdf2vec: Rdf graph embeddings and their applications. Semantic Web (Preprint), 1–32 (2018)
22. Ristoski, P., Schuhmacher, M., Paulheim, H.: Using graph metrics for linked open data enabled recommender systems. In: International Conference on Electronic Commerce and Web Technologies. pp. 30–41. Springer (2015)
23. Ristoski, P., de Vries, G.K.D., Paulheim, H.: A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In: International Semantic Web Conference. pp. 186–194. Springer (2016)
24. Rosati, J., Ristoski, P., Di Noia, T., Leone, R.d., Paulheim, H.: Rdf graph embeddings for content-based recommender systems. In: CEUR workshop proceedings. vol. 1673, pp. 23–30. RWTH (2016)
25. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: International Semantic Web Conference. pp. 245–260. Springer (2014)
26. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
27. Thalhammer, A., Rettinger, A.: Pagerank on wikipedia: towards general importance scores for entities. In: European Semantic Web Conference. pp. 227–240. Springer (2016)
28. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
29. de Vries, G.K.D., de Rooij, S.: Substructure counting graph kernels for machine learning from rdf data. Web Semantics: Science, Services and Agents on the World Wide Web 35, 71–84 (2015)
30. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: Twenty-Eighth AAAI conference on artificial intelligence (2014)