Towards a Benchmark for LOD-Enhanced Knowledge Discovery from Structured Data

Jindřich Mynarz and Vojtěch Svátek

Department of Information and Knowledge Engineering,
University of Economics, W. Churchill Sq.4, 130 67 Prague 3, Czech Republic
{jindrich.mynarz|svatek}@vse.cz

Abstract. To leverage on KDD architectures designed for business databases, original unbounded RDF data has to be transformed. We report on a use case consisting in adaptation of linked data, around a nucleus of public procurement data, for a data mining challenge event. The generic problems addressed are: linked data sampling; (generalised) concise bounded description extraction; propositionalisation to CSV using SPARQL `SELECT`; and aggregation behaviour assigned by checking conformance to eligibility criteria formulated as SPARQL queries.

1 Introduction

The discipline of Knowledge Discovery in Databases (hereafter KDD) emerged in early 1990s, and was characterized, among other, as a *"non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data"* [1]. In comparison with other related notions such as 'inductive inference' or 'machine learning', it typically has a business connotation, which is reflected by the 'usefulness' and 'comprehensibility' attributes in the definition. Classical KDD focuses on regularly structured data stored in (mostly relational) databases. In contrast, 'knowledge extraction' from the web of linked data, which emerged more than a decade later, usually focuses on natively graph-structured data often derived from free or semi-structured text (such as that of Wikipedia). Moreover, in case of RDF, the boundaries between data mining and querying are particularly blurry. Since most of the low-level knowledge structures are already materialized in the semantic description of data, data mining needs to concentrate on discovering higher level patterns in data. With this focus, data mining methods applicable on linked data (LD) typically have to be either designed from scratch or derived from pre-existing but specifically flavoured mining methods (text mining, graph mining etc.). The KDD mainstream thus remains isolated from the usage of the Linked Open Data (LOD) cloud or even LD in general.

In order to leverage on the power of scalable KDD architectures originally designed for business (or, e.g., public sector) databases, original 'unbounded' RDF data has to be transformed to a suitable shape and format. We report on an ongoing process of such transformation, aiming at designing a benchmark for mainstream KDD technologies that could benefit from available LD. We specifically elaborate on two aspects of the transformation: data *sampling* (extraction of a manageable RDF graph from the large set of interconnected data) and data *propositionalization* (which brings the data within the reach of tools that cannot handle multiple interlinked tables at a time), which includes data *regularization* (making it expressible in a concise form in a conventional format such as CSV). The transformation attempts to explore some new paths, such as a *generalized* form of the so-called *concise bounded description* in RDF extraction; it should be however noted that this kind of technological contribution is so far unproven to overcome pre-existing research; therefore, an (at least) equally important aspect of the research contribution is the nature of the core dataset used: *public procurement* data being substantially different from traditional, mostly encyclopaedic, linked data sources addressed in semantic data mining projects.

The paper is structured as follows. Section 2 explains the nature of the core dataset. Section 3 and 4 are devoted to the two crucial phases of the construction of the benchmark dataset: sampling and propositionalization. Finally, Section 5 surveys some related research and Section 6 wraps up the paper.

2 Domain and Context

The public procurement domain is fraught with numerous opportunities to corruption, while also offering a great potential for cost savings through increased efficiency. For example, it is estimated that the public procurement market accounts for 17,3 % of EU's GDP (as of 2008) [12], hence optimization in this area, including detection of fraud and manipulative practices, truly matters.

For this reason we started to build a benchmark dataset, primarily for the sake of a prospective first edition of a Linked Data Mining Challenge (LDMC) to take place this year.[1] The challenge will feature

1. A *descriptive task*, aiming at hypotheses to be further investigated by domain experts;
2. Two *predictive tasks*, specifically,
   - prediction of the *number of tenders* submitted for a particular call
   - classification of a public contract as *multi-contract* (conjoining goods/services of dissimilar nature).

The datasets prepared for the tasks of LDMC will consist of UK+US public procurement data, interlinked to data from the Linked Open Data Cloud[2] such as DBpedia.[3]

3 RDF Data Sampling

The first part of preparation of a benchmark dataset for data mining on LD consists in extracting a sample. A key question that arises when sampling linked

[1] As part of the DMoLD'13 workshop, see `http://keg.vse.cz/dmold2013/`.

[2] `http://lod-cloud.net/`

[3] `http://dbpedia.org/`

data is the definition of resource representation. Due to the unbounded nature of LD resources, there is no single or straightforward solution for determining the boundaries of resource representation that would fit all purposes. For example, representations meant for *user interfaces* commonly include resource *labels*. Our goal however was to provide a resource representation suited for KDD. For this purpose of which we came up with a generalised version of the established notion of *concise bounded description* (CBD). CBD [13] defines the scope of resource representation as including the *outbound* triples, i.e., those having the resource in the subject position, while for every such triple having either a blank node or an instance of `rdf:Statement` in its object position the CBD of its object is included recursively. There are several CBD variations such as the symmetric concise bounded description (SCBD) that adds *inbound* triples for which the described resource is used in the object position. The specification of CBD acknowledges that the description may span over multiple resources, however it does so in a rudimentary way by overloading the semantics of blank nodes and reified triples to delimit the boundaries of the description.

For many practical tasks a higher-level view of 'entities', rather than that of individual resources an entity may be composed of, is more appropriate. Hausenblas [3] defines an entity as *"a thematic view on resources across connectors, materialised through hyperlinks"* while *"data belonging to an entity is potentially distributed over several data sources."* This approach is in line with the SPARQL 1.1 Specification [2], which notes that the results of calling the SPARQL `DESCRIBE` query form, traditionally implemented as some form of CBD, *"may include information about other resources."*

To reflect this intuition we propose the *generalised concise bounded description* (GCBD), which extends the scope of CBD to cover all resources that are marked up as 'dependent' on the described resource (entity). GCBD is a representation of a resource that includes both outbound triples and inbound triples, but adds such triples recursively for all *dependent* resources, i.e., resources that carry data specific to an entity. This includes both resources identified with blank nodes and instances of classes a priori annotated as dependent. The dependency can be detected either at syntactical level, e.g., for triples reified using `rdf:Statement`, or at semantic level, e.g., for reified n-ary relationships identified in a manual or semi-automatic manner, for example, using the PURO ontological backrgound model [14] (where the annotation is assumed to be carried out at the level of vocabulary using a Protégé plugin). The choice of classes to annotate as dependent thus proceeds from a particular use case; it may include a common core, such as classes from the RDF namespace, and a use-case-specific part containing annotations of domain ontologies and vocabularies.

GCBD can be implemented as a SPARQL `CONSTRUCT` query with a transitive subquery that expands the description through dependent resources. Such *class annotations* are loaded into a separate named graph that is then referenced in the sampling queries. A straightforward way would be to attach annotations directly to RDF classes; however, since the type of instances may not be present

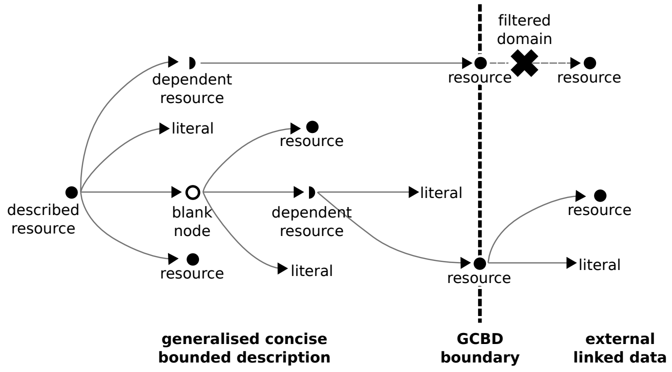

Fig. 1: Generalised concise bounded description and linked data

in source data (and we do not want to materialize it via `rdfs:range` inference) we opted in for attaching the annotations through properties.

Sampling the resource representations for the benchmark dataset is then straightforward. The sampler is given a SPARQL endpoint URL, a SPARQL `SELECT` query template to retrieve the resources of interest, a list of URIs of named graphs to sample data from, a target named graph URI, access credentials to allow SPARQL Update operations, the sample size and the preferred language tag to be used for sampling the labels. After checking if any data matches the provided requirements, the sampler extracts a randomly selected subset of resources. Consequently, for each resource in the subset its entity description is inserted into the target named graph via a SPARQL Update request.

Since we are dealing with linked data, the sample data retrieved via a SPARQL endpoint may be combined with data available through resolving linked URIs of resources from *external datasets*. In the course of sampling, we only collect links already present in linksets created by executing Silk[4] linkage rules inside the ODCleanStore framework.[5] We issue a SPARQL query that computes an intersection of the sample data with relevant linksets and produces a list of external URIs, which may be optionally narrowed down to include only URIs from domains of interest. Linked resources on the list are harvested using LDSpider,[6] performing a breadth-first crawl, and loaded into the sample named graph. The extent of GCBD and associated linked data is depicted in figure 1.

[4] `http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/`

[5] `http://sourceforge.net/p/odcleanstore`

[6] `https://code.google.com/p/ldspider/`

4 Data Propositionalization

Propositionalization in KDD means transforming relational data into a single table with a set of propositions in the form of attribute-value pairs [7]; some structural information is thus lost in aggregations and some relationships in data discarded. Owing to the malleable nature of RDF and flexibility of SPARQL, linked data can be propositionalized via the SPARQL `SELECT` query form, allowing to transform graph-shaped data into the tabular format [4], which can then be exported in the comma-separated values (CSV)[7] data format. Such a query, containing a `GROUP BY` clause on variables binding the described resource, can be built programmatically based on the dataset structure previously revealed by exploratory SPARQL queries. For each property a subquery is created that specifies the handling of its object values and their *aggregation behaviour* according to eligibility criteria (formalized as SPARQL queries, too).

The conformance of a property with the criteria is schema-agnostic: rather than from the property definition in its vocabulary, it is empirically inferred from the way the property is used in the processed dataset, primarily from its *cardinality* (assuming it is higher than one). Basic aggregations may be accomplished with in-built SPARQL aggregates, including `COUNT`, `SUM`, `MIN`, `MAX`, `AVG` and `GROUP_CONCAT`. SPARQL `SAMPLE` selects a random binding in a non-deterministic fashion, which makes it inappropriate for our purposes. With the exception of `COUNT` and `GROUP_CONCAT`, SPARQL aggregations are intended to be applied primarily on numeric values, therefore the eligibility criterion for such aggregations contains a `FILTER` with the `isNumeric` function. A simple measure to fold multiple values into one is to choose a ‘default’ or ‘preferred’ value (with the SPARQL `IF` functional form or `COALESCE` function). For example, in multilingual datasets such approach may be used to pick a literal value with a preferred language tag, in the case of labels.

While the rich structure of LD makes automation of aggregation difficult, it can be, on the other hand, exploited for ‘smart’ aggregations. For instance, generalization to a common broader concept may be applied if the property range values come from hierarchically structured taxonomies or partonomies (e.g., `skos:ConceptSchemes`) [8]. Resource descriptions can also be expanded through the linked resources’ URI references, since if a property has a non-literal range, the aggregation behaviour may be recursively applied to its object values; for example, n-ary relationships can be decomposed into multiple columns. A question is then how the graph traversal depth should be set, especially when it comes to traversing linked external resources. In our case, the traversal boundaries are established by the scope of the sample (as defined using GCBD) or can be set arbitrarily when crawling the linked resources. Properties could also be equipped with manually selected aggregation behaviour expressed in the SPIN SPARQL syntax,[8] which supports basic SPARQL aggregations and allows to express arbitrary queries.This approach opens an opportunity for further research

[7] `http://tools.ietf.org/html/rfc4180`

[8] `http://spinrdf.org/sp.html`

on the use of machine learning to assign the aggregation behaviour automatically. Another challenge requiring further research is to specify how to handle properties with heterogenous ranges (e.g., mixing literal and URI references).

5 Related Research

The related research can be divided into data mining approaches applied to public procurement data on the (semantic) web and data mining on semantic web data in general.

Even though there seems to be much to gain by applying data mining methods in the procurement domain in connection with LD, relatively few projects explored it. One such undertaking is the Linked Open Tenders Electronic (LOTED) project [15]. LOTED triplified data from RSS feeds published by Tenders Electronic Daily, an European Commission portal aggregating public procurement data from the EU member states. After converting the data to RDF and linking it, LOTED was able to apply lightweight data analysis methods and showcase interactive visualizations. Furthermore, Monteiro [9] applied data mining for outlier detection to spot fraud in Portuguese public procurement. However, in this case the data used were structured in XML and it took into account only a limited subset of data including contract description and award price.

In terms of applying semantic data in data mining, Liu [8] provided a general outline for incorporating semantic technologies in data mining, such as with semantic annotation. Kiefer et al. [6] proposed SPARQL-ML, an approach to data mining on semantic web data focused on statistical relational learning and SPARQL. Similar direction is followed by the RMonto tool [11], an upper layer for the popular RapidMiner tool. It allows to apply ontologies as background knowledge for several mining tasks, possibly combining relational and propositional subtasks. Another implementation of RDF data pre-processing for RapidMiner [5]. Finally, Paulheim & Fürnkranz [10] suggested an automated method for data enrichment from Linked Data, pipelining entity recognition, feature generation and feature selection. These generic projects are, unarguably, technologically more mature than our effort, primarily launched as rapid and self-governed means for building a concrete challenge dataset; we thus assume that such systems are likely to be directly applicable on the RDF modality of our procurement data and would achieve propositionalization by themselves if participating in the Linked Data Mining Challenge. Our effort is to a large degree orthogonal to the research of semantic web data mining algorithms proper; it strives to build a single resource containing the same data objects in multiple levels of complexity, thus potentially allowing relational DM tools to compete with propositional ones. A relatively well structured real-world dataset was chosen so as not to disqualify the non-relational (but highly scalable) tools from the beginning.

6 Conclusions

In general, there are two ways to facilitate data mining on linked data. First, transform data to a form more familiar to the existing data mining tools, thus lowering the barrier to applying established data mining methods. Second, to move the KDD tools closer to linked data, so that they can work with it natively. Although we presume that a more attractive opportunity lies, in long term, with data mining tools exploiting the rich structures of graph-shaped data (e.g., RDF) directly, we selected to explore the first approach. While the transformation of linked data to tabular format entails serious information loss, this approach has at least two positive aspects: 1) efficient processing of data that is already regularly structured (such as the core LDMC dataset harvested from the centralized public contract servers) is preserved, and, 2) business analysts can interact with familiar tools and thus fully exploit their competence.

In the above-described work we experimented with transforming highly-relational linked data into a propositional dataset. Such transformation is, necessarily, a lossy process that 'downgrades' RDF into a tabular format, while giving up some of the structural information and materialized relationships. This approach is a subject to trade-off between the extent of data loss and ease of use with existing data mining software. Consequently, our further work needs to validate whether such approach yields data that is easily amenable to data mining tools or if it produces data crippled to such an extent that few insights may be discovered in it. We also plan to compare the added value of GCBD compared to CBD, applied in the extraction phase, in terms of the mining result quality.

Acknowledgment

The research has been partially supported by the EU ICT FP7 under No. 257943, LOD2 project.

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (eds.) (1996): *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press.
2. Harris, S., Seaborne, A. (eds.): *SPARQL 1.1 Query Language.* W3C Proposed Recommendation 08 November 2012. Online: `http://www.w3.org/TR/sparql11-query/`.
3. Hausenblas, M. (2011): On Entities in the Web of Data. In Wilde, E., Pautasso, C. (eds.). *REST: From Research to Practice.* ISBN 978-1-4419-8302-2.
4. Hausenblas, M., Villazón-Terrazas, B., Cyganiak, R. (2012): *Data Shapes and Data Transformations.* CoRR. Online: `http://arxiv.org/abs/1211.1565`.
5. Khan, M.A., Grimnes, G.A., Dengel, A. (2010): Two pre-processing operators for improved learning from SemanticWeb data. In: RapidMiner Community Meeting and Conference: RCOMM 2010 proceedings.
6. Kiefer, C., Bernstein, A., Locher, A. (2008): Adding data mining support to SPARQL via statistical relational learning methods. In: Proceedings of the 5th European semantic web conference (ESWC'08), Springer-Verlag, Berlin, Heidelberg, 478-492.

7. Lachiche, N. (2013): Propositionalization. In *Encyclopedia of Machine Learning*. Springer.
8. Liu, H. (2010): Towards Semantic Data Mining. In *ISWC'2010*. Online: `http://ix.cs.uoregon.edu/~ahoyleo/research/paper/iswc2010.pdf`.
9. Monteiro, J. A. (2011): The Guardian of the Republic: A conceptual system to detect outliers on Public Contracts. In *Proceedings of the 6th Doctoral Symposium on Informatics Engineering*. Online: `http://paginas.fe.up.pt/~prodei/dsie11/images/pdfs/s1-2.pdf`.
10. Paulheim, H., Fuernkranz, J. (2012): Unsupervised generation of data mining features from linked open data. In: International Conference on Web Intelligence, Mining, and Semantics (WIMS12).
11. Potoniec, J, Lawrynowicz, A. (2011): RMonto: Ontological extension to RapidMiner. In: Poster and Demo Session of the ISWC 2011 – 10th International Semantic Web Conference, Bonn, Germany.
12. *Study on the evaluation of the Action Plan for the implementation of the legal framework for electronic procurement (Phase II): Analysis, assessment and recommendations*. Version 3.2. 9 July 2010. Online: `http://ec.europa.eu/internal_market/consultations/docs/2010/e-procurement/siemens-study_en.pdf`.
13. Stickler, P.: *CBD - Concise Bounded Description*. W3C Member Submission 3 June 2005. Online: `http://www.w3.org/Submission/CBD/`.
14. Svátek V., Homola M., Kluka J., Vacura M. (2013): Metamodeling-Based Coherence Checking of OWL Vocabulary Background Models. In: Proc. OWLED'13, Montpellier, to appear.
15. Valle, F., d'Aquin, M., Di Noia, T., Motta, E. (2010): LOTED: Exploiting Linked Data in Analyzing European Procurement Notices. In *KIELD'2010*. Online: `http://ceur-ws.org/Vol-631/paper6.pdf`.