Linked Data Mining Challenge (LDMC) 2013 Summary

Vojtěch Svátek, Jindřich Mynarz, and Petr Berka

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

Abstract. The paper summarizes the conception, data preparation and result evaluation of the LDMC, which has been organized in connection with the DMoLD'13 - Data Mining on Linked Data Workshop, Prague, September 23 (as part of the ECML/PKDD conference program).

1 Introduction

The organization of contests or 'challenges' has long tradition both in data mining (KDD Cup, ECML/PKDD Discovery Challenge, Kaggle.com, and many other, see also `http://www.kdnuggets.com/competitions/`) and semantic web (SW Challenge, LinkedUp Challenge, USEWOD Challenge, OAEI, etc.) However, the intersection of these two fields doesn't seem to sufficiently exploit the potential of open competitions yet. USEWOD Challenge and OAEI definitely have aspects of 'data mining', however, they focus on mining problems specific for the semantic web field itself (linked data usage analysis and ontology matching, respectively). What has been missing was a challenge event addressing a real *business knowledge discovery problem*, for which semantic web approaches would be beneficial, be it thanks to their data modelling flexibility structure or thanks to their capability of interlinking data from independent, heterogeneous sources. This event should give priority to reuse and adaptation of 'the best of the breed' from the long KDD tradition rather than to inventing linked-data-tailored approaches from scratch; in this respect it should be tied to a data mining rather than core semantic web event.

The Linked Data Mining Challenge (LDMC) has thus been envisaged to fill this important gap, and, generally, to spur the research collaboration between the semantic web community (represented by the linked data sub-community as its practice-oriented segment) and the data mining community.

This summary paper describes the conception, data preparation and result evaluation for the first LDMC edition, organized in connection with the DMoLD'13 - Data Mining on Linked Data Workshop, in Prague on September 23, as part of the program of ECML/PKDD, one of the most recognized scientific conferences in the data mining field.

The paper is structured as follows.

2 Business Domain, Mining Tasks and Underlying Data

2.1 Business Domain

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) [10], hence optimization in this area, including detection of fraud and manipulative practices, truly matters.

Data from this domain are frequently analyzed by investigative journalists and transparency watchdog organizations; these, however,

1. rely on interactive tools such as OLAP and spreadsheets, incapable of spotting hidden patterns, and
2. only deal with isolated datasets, thus ignoring the potential of interlinking to external datasets.

Focusing (one or multiple editions of) the LDMC activity on this domain could possibly initiate a paradigm shift in analytical processing of this kind of data, eventually leading to large-scale benefits to the citizenship.

2.2 Mining Task Formulation

Due to the novelty of this kind of challenge, it was originally assumed that the task would be merely *exploratory*: mining for interesting hypotheses of any kind, whose plausibility and novelty would be subsequently judged by domain experts. However, the conjunction of the challenge with the ECML/PKDD conference, where predictive data mining predominates, made us also start considering suitable *predictive* tasks. Eventually, the first, 2013, edition of LDMC has been set up as a combination of three tasks, of which Tasks 1 and 2 would be predictive and Task 3 would amount to 'free exploration'.

In all tasks the experimenters were assumed to make use of linked data resources. Some external resources have been already interlinked to the original dataset. It was also possible to heuristically link further resources from the Linked Data Cloud. The data provided was not fully cleaned and it did not adhere to the used ontologies entirely, especially regarding cardinalities.

2.3 Task 1

The task was to predict the *number of bidders* (as integer value). In the training dataset, the number of bidders was expressed as value of the *pc:numberOfTenders* property. The preciseness mattered most for the lower values, e.g., predicting 2 bidders where there are 3 is a more important error than predicting 12 bidders where there are 13. This has been reflected in the evaluation measure.

The dataset was divided as follows:

– The training dataset contained 1,658 US public contracts (as instances of *pc:Contract class*), with the number of submitted tenders known. There has been 38,743 RDF triples, and 469 *owl:sameAs* links with external entities.
– The testing dataset contained 1,737 notices of US public contracts (as instances of *pc:Contract class*) that were still open to bidders. There have been 37,489 RDF triples, and 346 *owl:sameAs* links with external entities. The number of tenders (i.e. *pc:numberOfTenders*) for the public contracts was not yet known and thus not featured in the data. The selected contracts were assumed to be closed shortly after the deadline for submitting results for this task.

The results for the task were required to be delivered in CSV format with two columns. First column would contain the URI of an annotated public contract, and the second column would contain the predicted number of tenders for the public contract in the format of positive integer.

The principal evaluation measure at the level of individual object has been the absolute value of the difference between the predicted value $\bar{v}$ and the reference value v, adjusted by the reciprocal value of the (smaller, except zero) value size and normalized to [0,1] by a sigmoidal function:

$$Err(v, \bar{v}) = \frac{2}{1 + e^{-\frac{\|v - \bar{v}\|}{max(1, min(v, \bar{v}))}}} - 1$$

The adjustment by reciprocal value made the cost of errors uneven for the same value difference (same difference for larger values counting less than that for smaller values). The error values were to be aggregated by average.

2.4 Task 2

The task was to classify the contracts as *multi-contract* or its opposite. A multi-contract is a contract that (often, 'suspiciously') unifies two or more unrelated commodities. It is also possible to classify a contract as borderline. In the training dataset, the multi-contract annotation is expressed as value of the artificially added *multicontract* property.

Due to difficulties in the manual annotation process, the datasets were rather tiny:

– The training dataset contained 40 multi-contracts and 168 non-multi-contracts. There have been 141,976 triples, and 372 *owl:sameAs* links to entities.
– The testing dataset contained 10 multi-contracts and 42 non-multi-contracts (to keep the positive/negative ratio equal as in the training set). There have been 60,518 triples, and 82 *owl:sameAs* links to entities.

The data corresponded to UK public contracts, plus CPV codes and DBpedia entities.

The results for the task were be delivered in CSV format with two columns. The first column should contain the URI of an annotated public contract, and

the second column should contain annotation for the predicted variable with three possible values: 0 if the contract is not a multi-contract, 0.5 if the contract is a borderline case, and 1 if the contract is a multi-contract.

The evaluation measures considered have been:

- *Accuracy:* average distance between the predicted values and the reference values, all of which can take the discrete values from the set 0, 0.5, 1, see above.
- *Precision:* the proportion of predicted multi-contracts (i.e. predicted value 1) that are indeed multi-contracts (i.e. reference value 1).
- *Recall:* the proportion of true multi-contracts (i.e. reference value 1) that are predicted as multi-contracts (i.e. predicted value 1).

2.5 Task 3

The task was to find (and possibly attempt to suggest explanations to) any kind of interesting hypotheses (nuggets) in data. An example could be hypotheses related to uneven distributions of CPV codes in different geographical segments of contracts data, but many other options were possible.

The data contained 5,002 instances of *pc:Contract*, described using 431,300 RDF triples; there have been also 5,120 *owl:sameAs* links The contract data came from the following resources

- `https://www.fbo.gov/`
- `http://usaspending.gov`
- `http://contractsfinder.businesslink.gov.uk`
- `http://linked.opendata.cz/resource/dataset/far-codes` (FAR codes referred to by the US contracts, as collected by the LDMC team)
- `http://linked.opendata.cz/resource/dataset/cpv-2008` (CPV codes referred to by the UK contracts, as collected by the LDMC team)
- `http://dbpedia.org/` (encyclopaedic data)

Evaluation was supposed to based on interestingness of the findings (and possibly their interpretation), described in a submitted paper and judged by experts in public procurement. Although Task 3 was possibly more relevant for getting practical insights into the data from the business point of view, it has not attracted enough attention (possibly due to ECML/PKDD bias towards computational rather than business aspects of KDD) and has not been addressed by any of the submissions.

2.6 Vocabularies Used

The data for all three tasks has been modelled using RDF vocabularies and ontologies, including the following:

- Public Contracts Ontology (`http://purl.org/procurement/public-contracts#`)
- Schema.org (`http://schema.org/`)

– GoodRelations (`http://purl.org/goodrelations/v1#`)
– Dublin Core Terms (`http://dublincore.org/documents/dcmi-terms/`)
– Simple Knowledge Organization System (`http://www.w3.org/TR/skos-reference/`)
– VCard (`http://www.w3.org/2006/vcard/ns#`)
– Asset Description Metadata Schema (`http://www.w3.org/ns/adms#`)

The format of data was RDF in Turtle serialization (GZipped).

3 Data Selection and Preparation Process

There has been very little previous experience with preparing RDF data for analysis by mainstream data mining tools. The process of data preparation proved more difficult than initially expected. By consequence, both Task 1 and Task 2 suffered from the extremely small size of the samples. We will now discuss problems encountered when preparing the datasets for both *predictive tasks*.

3.1 Data Selection and Extraction

Data for *Task 1* depended on the availability of the number of contract bidders in source data. In many public procurement datasets this number is currently not disclosed, for example in the case of British *ContractsFinder*. Therefore, our choice of datasets from which to draw the sample for this task was severely limited. Due to our limitation to English language data, we chose to use data from *USASpending.gov*, which provides data on the number of business entities that bid on contract notices. Having chosen this source soon we discovered that it only provides data about awarded public contracts, all of which already have number of bidders published. While the retrospective disclosure was sufficient for the training sample, it was not possible to prepare a testing sample from such data. For the purposes of testing dataset we hoped to obtain a selection of public contracts notices, for which the deadline for submitting tenders was situated in the future, so that the number of bidders would be unknown prior to the delivery of LDMC's results. In such a selection the number of bidders would be revealed after the deadline for submitting results to LDMC, typically in contract award notices. However, since all public contracts in the *USASpending.gov* dataset had already been awarded, we needed to find another data source, which publishes contract notices. While there are several such sources, the major one that we found is *FedBizOpps.gov*, which conveniently provides a data dump in XML, thanks to the *Data.gov* (U.S.) open data initiative.

We transformed the CSV data from *USASpending.gov* to RDF using a SPARQL mapping executed by *tarql* (`https://github.com/cygri/tarql`). Similarly, we converted the *FedBizOpps.gov* dataset from XML to RDF using a custom XSLT stylesheet. Having both datasets in RDF, we needed to establish identity links between the same public contracts from the two sources. We used a simple *Silk* (`http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/`) linkage rule based on contract identifiers and associated version numbers, with which we however only managed to link a small fraction of public contracts present in both

dataset. The low recall of the linkage rule was caused by differences in identifiers used for public contracts in the processed datasets. Even though some differences have been smoothed by applying a straightforward normalization, we had to omit public contracts with widely distinct, erroneous (such as the omnipresent "0001" code) or even missing identifiers. Ultimately, we obtained 1658 identity links between public contracts in the two datasets, all of which were used for the *training* sample for Task 1.

To prepare the *testing* dataset for Task 1 we needed to collect public contracts for which the number of bidders would be disclosed during the interval between the deadline for LDMC's results submission and the publication date of LDMC evaluation results (minus a few days to allow for data processing and evaluation). Based on an analysis of the typical delay between the deadline for tender submission and the date when the number of bidders is published, we opted for a delay of 1 month, meaning that the testing dataset included public contracts notices with tender submission deadline from the interval starting a month before the LDMC's results submission deadline to at least a month prior to the publication date of LDMC evaluation results. Such selection criteria yielded 1737 public contracts. Knowing about the lossy nature of linking these 2 sources and given the assumption that for a large part of the contracts award the data is not published at all, we expected to be able to obtain the actual number of bidders only for a fraction of the testing dataset. During the submissions' evaluation we harvested recent data from *USASpending.gov* and interlinked it with data from the testing dataset using the same procedure as described above. This exercise yielded 50 public contracts for which we got the actual number of bidders. Consequently, the evaluation of task 1 was based on this fractional subset of the original testing data.

Preparation of datasets for Task 2 faced different challenges. Because of our restriction to English language data, we chose to use the British *Contracts-Finder* application as a source, since it provides XML exports, which we had already converted into RDF. However, as we learnt afterwards, there were several problems with missing data in the converted output, which were due to the overly imperative nature of the XSLT stylesheet that had been used to execute the transformation. For example, additional Common Procurement Vocabulary codes were lacking at times, since the stylesheet had been written so as to expect exactly one additional code. The crucial role in preparation of the datasets for Task 2 was enacted by two *domain experts* whom we contracted to annotate the data. Their task was to classify the data either as multi-contracts or as non-multi-contracts. We coined the term 'multi-contract' to denote public contracts that bundle unrelated products or services (for example, software and cleaning services), which in fact should be split into multiple separate contracts. In order to pre-filter the public contracts for manual annotation, we extracted public contracts that had the least similar main object and one of their additional objects. The objects of contracts in British *ContractsFinder* are expressed using the Common Procurement Vocabulary (CPV), a standard code list for public procurement in the EU. CPV has a hierarchical structure, so we inferred the

dissimilarity of two CPV codes from their distance in the hierarchical tree of CPV. Having the CPV data previously converted to RDF, we merged it with British contracts data and extracted contracts with 1000 least similar codes. The resulting list of contract URIs was transformed into a readable table using a complex hand-crafted SPARQL SELECT query. In this way, we managed to provide the domain experts with a friendlier representation than would be raw RDF, while also presenting them with only a manageable subset of contracts that were more likely to be annotated as multi-contracts. Unfortunately, due to the prolonged preparation time and limited availability of domain experts we ended up with 200 annotated public contracts only. Moreover, as we learnt from the domain experts, pre-filtering public contracts based on distance in the CPV hierarchy did not work as expected because of the existence of closely related CPV codes contained in completely different branches of the code list.

The annotated public contracts have been split into two parts, the first of which was published as the training set, while the second was stripped from annotations and used as the testing set. The annotations for the testing set were withheld in order to serve as validation data.

Preparation of data for *Task 3* was by far the least problematic. Much of its ease of preparation can be ascribed to minimal requirements on its output. Since Task 3 was about open exploration of data, it was possible to provide just the available data without much preprocessing. Having transformed both source datasets for Tasks 1 and 2, i.e. *USASpending.gov* and British only *Contracts-Finder*, we selected 5,002 random instances of public contracts coming from a mix of the two sources, which were then used for the dataset of Task 3.

All datasets prepared for the LDMC had been subject to basic cleaning and deduplication. On the other hand, at this step the datasets had been polluted by artefacts and by-products of cleaning procedures that left traces of their materialized auxiliary data in their output.[1] Datasets had been enriched by geocoding postal addresses present in them. Business entities participating in public procurement, i.e. contracting authorities and bidders, had been linked to corresponding DBPedia resources. However, due to poor quality of the data, lacking strong identifiers, the precision and recall of the external linkage was very low. Finally, datasets had been enhanced with DBPedia data harvested by following their links by LDSpider (`https://code.google.com/p/ldspider/`), which was configured to fetch resources within the 1-hop neighbourhood. Unsurprisingly, supplying potentially relevant data from DBPedia came at the cost of further increasing the level of noise in the resulting sample datasets.

3.2 Data Propositionalization

In order to further lower the barrier of entry for LDMC participants, we originally planned to deliver the datasets for all tasks in the form of relational tables or a

[1] In particular, apart from pruning identical resources, the deduplication also led to over 14 thousand superfluous *owl:sameAs* links that connected resources with the same URIs, due to a bug in the version of Silk we used.

single table in CSV. The reasoning behind this decision was motivated by the recognition that most existing data mining tools are not capable of handling RDF, whereas they support well tabular data either in relational tables or in CSV. The same motivation drove Ramanujam et al. [7] when they proposed transforming RDF into relational structures to enable reuse by existing tools. An additional advantage of having data in CSV was the possibility to use Kaggle[2] competitions as parallel platform for undertaking analysis of the same data.

Unfortunately, in the end we did not manage to provide propositionalized data to LDMC participants. The main cause behind this state of affairs was due to our underestimation of time and resources needed to develop a solution for RDF propositionalization. The primary source of complexity was our decision to transform RDF in a schema-agnostic fashion, so that the mechanism is not task-specific and is able to recursively process previously unknown linked data. In order to fulfil this requirement, we had to program automatic discovery of empirical schema of data via exploratory SPARQL queries. An additional source of complexity arose from the messiness of processed data, which had to be normalized and enriched before proceeding with propositionalization.

Our intention was to deliver LDMC datasets in two additional forms: *set of relational tables* and *single aggregated table*. When the goal is to produce a single table from RDF, the most naive implementation could use a table with 3 columns for subjects, predicates and objects, which is not all that usable. A slightly more sophisticated approach is to convert RDF into a set of per-property tables, in which each property becomes the name of a table containing 2 columns for subjects and objects associated with the property. To get closer to the typical form of relational data Ramanujam et al. [7] proposed to transform RDF into tables per each class, containing all properties used with the class instances in 0...1 cardinality, while for each property of higher cardinality a separate table is created. To execute the transformation from RDF to tabular data, SPARQL SELECT query form is a well-suited option [1]. Thanks to its strong SQL heritage, SPARQL allows to implement most aggregation functions used for propositionalization in SQL [8]. Moreover, SPARQL 1.1 provides the results directly in standardized tabular format, either in CSV or TSV [9]. The details of our initial plan for RDF propositionalization are described in our previous paper [6].

4 Summary of Participants' Results

The response to the challenge call was relatively weak, presumably due to the novelty of the task and small number of researchers with sufficient expertise in both data mining and linked data. Task 1 has been addressed by two groups, from the Technical University of Darmstadt (TUD) and from the Vrije Universiteit Amsterdam (VUA). Task 2 has only been addressed by the (same) VUA group. Task 3 (free exploration) has unfortunately not attracted any participant.

[2] `http://www.kaggle.com`

4.1 Task 1

Both participants obtained comparable values of the error formula on the validation set. TUD was slightly better (0.3747) than VUA (0.3849). Both results are worse than a constant-value predictor with the constant set to any value between 4 and 7 (for VUA also including 3). The best constant-value predictor was obtained for the most frequent value in the validation set, 4, with $Err(v, \bar{v}) = 0.3057$. However, for value 1, which was most frequent in the training set, the constant predictor would perform much worse (0.8138). The disbalance of class values in the training and validation datasets (possibly related to the temporal shift: the validation dataset contained newer contracts due to requirement of their unknown result at the time of participants' analysis) was probably caused by the overall small size of the data: only 50 examples have been eventually used for validation.

4.2 Task 2

The only participant, VUA, reached the accuracy of 0.7885. Similarly to Task 2, this result is below the baseline corresponding to predicting the most frequent value (non-multi-contract), which is 0.8077. The precision was 0.3333 (out of the 3 predicted multi-contracts, one was a labelled as such) and the recall was 0.1 (one of the ten labelled multi-contracts was predicted as such). Somewhat surprisingly, the experimenters did not seem to exploit the information about the overall number of multi-contracts in the validation dataset, which was publicly available on the website.

4.3 Domain Experts' View

The authors of LDMC submissions avoided interpretation of their results and instead focused on technical aspects of the data mining techniques employed. No submission was received for the open, exploratory task 3. Therefore, domain experts were not able to judge the relevance of submissions.

4.4 Syntactical Issues of Results Delivered

Instead of URIs of contracts, the TUD team used URIs of their identifiers (values of *adms:Identifier*). Furthermore, all 3 delivered submissions were not formally valid CSV.

5 Conclusions

The first edition of LDMC already provided useful feedback to its organizers. Discussion to be held at the DMoLD'13 workshop is likely to led to improvements in the challenge's setting. A crucial question to be solved is why the impact of *external linked data* was claimed to be negligible for the result of predictive tasks.

In longer term, we plan to eventually complement RDF data by data transformed to the *CSV format*, including a single ('propositional') table. This will allow to address predictive tasks via the Kaggle platform. Regarding the collocation of the DMoLD workshop with the ECML/PKDD conference, it is to be determined whether this kind of conference is an optimal venue; a possible alternative would be a more business-oriented conference with focus on business aspects of data mining.

Acknowledgment

The preparation of LDMC and of this paper has been partially supported by the EU ICT FP7 under No. 257943, LOD2 project. The authors would like to thank Jakub Stárka for his involvement in the data extraction process and to the domain experts Jiří Skuhrovec and Jana Chvalkovská for general feedback.

References

1. Hausenblas, M., Villazón-Terrazas, B., Cyganiak, R. (2012): *Data Shapes and Data Transformations*. CoRR. Online: `http://arxiv.org/abs/1211.1565`.
2. 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.
3. 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.
4. Lachiche, N. (2013): Propositionalization. In *Encyclopedia of Machine Learning*. Springer.
5. Liu, H. (2010): Towards Semantic Data Mining. In *ISWC'2010*. Online: `http://ix.cs.uoregon.edu/~ahoyleo/research/paper/iswc2010.pdf`.
6. Mynarz, J., Svátek, V.: Towards a benchmark for LOD-enhanced knowledge discovery from structured data. In: Second International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD'12). Available from WWW: `http://ceur-ws.org/Vol-992/paper6.pdf`. ISSN 1613-0073.
7. Ramanujam, S., Gupta, A., Khan, L., Seida, S., Thuraisingham, B.: Relationalizing RDF stores for tools reusability. In: Proceedings of the 18th international conference on World Wide Web. New York (NY): ACM, 2009, pp.1059–1060.
8. Reutermann, P., Pfahringer, B., Eibe, F.: A toolbox for learning from relational data with propositional and multi-instance learners. In: Proc. 17th Australian Joint Conference on Advances in Artificial Intelligence. Springer, 2004, pp. 1017–1023.
9. Seaborne, A. (ed.). SPARQL 1.1 Query results CSV and TSV formats [online]. W3C Recommendation 21 March 2013. Available from WWW: `http://www.w3.org/TR/sparql11-results-csv-tsv/`
10. *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`.