The Linked Data Mining Challenge 2014
Results and Experiences

Vojtěch Svátek[1], Jindřich Mynarz[1], and Heiko Paulheim[2]

[1] University of Economics
Department of Information and Knowledge Engineering
Prague, Czech Republic
{svatek,jindrich.mynarz}@vse.cz
[2] University of Mannheim, Germany
Research Group Data and Web Science
heiko@informatik.uni-mannheim.de

Abstract. The 2014 edition of the Linked Data Mining Challenge, conducted in conjunction with Know@LOD 2014, has been the third edition of this challenge. The underlying data came from two domains: public procurement, and researcher collaboration. Like in the previous year, when the challenge was held at the *Data Mining on Linked Data* workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2013), the response to the challenge appeared lower than expected, with only one solution submitted for the predictive task this year. We have tried to track the reasons for the continuously low participation in the challenge via a questionnaire survey, and principles have been distilled that could help organizers of future similar challenges.

1 The Linked Data Mining Challenge Overview

Linked data (LD) represents a novel type of data source that has been so far nearly untouched by advanced data mining methods. It breaks down many traditional assumptions on source data and thus represents a number of challenges:

- While the individual published datasets typically follow a relatively regular, relational-like (or hierarchical, in the case of taxonomic classification) structure, the presence of semantic links among them makes the resulting 'hyper-dataset' akin to general graph datasets. On the other hand, compared to graphs such as social networks, there is a larger variety of link types in the graph.
- The datasets have been published for entirely different purposes, such as statistical data publishing based on legal commitment of government bodies vs. publishing of encyclopedic data by internet volunteers vs. data sharing within a research community. This introduces further data modeling heterogeneity and uneven degree of completeness and reliability.

– The amount and diversity of resources as well as their link sets is steadily growing, which allows for inclusion of new linked datasets into the mining dataset nearly on the fly, at the same time, however, making the feature selection problem extremely hard.

The motivation for organizing the Linked Data Mining Challenge (LDMC) was twofold. First, it aimed to advertise the large quantities of linked data recently arising [1, 7] to a community which may have an interest in such diverse real-world datasets for testing machine learning and data mining systems and algorithms. Second, the data mining experience provided by challenge participants could foster an exchange on ideas and methods addressing the particularities of Linked Data mining.

The call for challenge contributions was sent to several relevant mailing lists from the semantic web, data mining, as well as more general area (e.g., ML-news@googlegroups.com, public-lod@w3.org, semantic-web@w3.org, DBworld). However, the response was unsatisfying throughout all three editions. In summary:

– In 2012 there was no challenge result submission, and the workshop as such only attracted 1 submission and had to be canceled.
– In 2013 there were 3 challenge result submissions [2, 4]. On the other hand, there were 5 regular paper submissions to the workshop, and, most notably, the workshop attracted a significant number of participants (over 40).
– In 2014 there was only 1 challenge result submission [3]. There were 10 paper submissions and 25 participants registered to the workshop; however, this time the workshop itself was not primarily proposed as framing for the challenge, but was a continuation of a previously started series.

Both 2013 and 2014 editions were used as a platform to discuss, with the participants, the problems and opportunities of such a challenge event. Furthermore, the 2014 edition was followed by a closed questionnaire survey (only targeting the registered Know@LOD'14 workshop participants) aiming at learning lessons from the LDMC organization endeavor.

In this paper, we describe the challenge tasks, the datasets used, the process of data preparation, and, finally, the results of the questionnaire survey.

2 Tasks and Datasets

The 2014 edition of the Linked Data Mining Challenge comprised three tasks, one predictive and two exploratory tasks.

2.1 Ordinal Prediction Task

The ordinal prediction task was prepared for all three editions, and always related to the procurement domain. The target attribute to be predicted was the *number of tenders* for the respective public contract; the true value of this target attribute was not known for the evaluation dataset before the bidding period has been closed (which was after the result submission deadline).

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.

2.2 Exploratory Tasks

Exploratory tasks were considered as most important and thus prepared for all three editions. While in 2012 and 2013 the exploration only addressed the procurement domain, a second domain was added this year, i.e., researcher collaboration.

2.3 Datasets

For the ordinal prediction task, we used U.S. procurement data from two interlinked resources: FBO[3] and USA Spending[4] (the process of their extraction and interlinking is described in the LOD2 project deliverable D9a.3.1 [9]). That dataset was also used for the first exploratory task. The second exploratory task used linked data on Australian research institutions, collected as described by Myers et al. [6].

All task descriptions and datasets can also be found online.[5]

3 Data preparation

Since the process of data preparation for the 2013 edition is described in detail in [8], we concentrate on the (only slightly modified) process applied for the 2014 edition.

[3] `https://www.fbo.gov/`

[4] `http://usaspending.gov`

[5] `http://knowalod2014.informatik.uni-mannheim.de/en/linked-data-mining-challenge/`

3.1 Training and testing data

The dataset was created by combining data from two principal sources, which provide complementary kinds of data. The two sources in question are USASpending.gov[6] and Federal Business Opportunities (FBO).[7] USASpending.gov offers a database of government expenditures, including awarded public contracts, for which it records, e.g., the aforementioned numbers of bidders. On the other hand, FBO publishes public notices for ongoing calls for tenders. Once public notice's deadline for tender submission passes, final number of bidders should be published along with other information about contract award in USASpending.gov. Unfortunately, these two sources do not publish enough data about public contracts to pair the equivalent instances reliably. While the same contract identifiers are used in some cases, most of the published contracts lacks identifying information necessary for deduplication. Combination of data from the two sources thus yields only a small subset of public contracts that could be merged provided they are equipped with strong identifiers, such as URIs.

USASpending.gov provides data downloads in several structured data formats, including CSV, TSV, XML and Atom. We used the CSV dumps, which we converted to RDF using SPARQL mapping[8] executed by tarql.[9] Data dump from FBO is available in XML as part of the Data.gov initiative.[10] To convert the data to RDF we created an XSLT stylesheet that outputs RDF/XML.[11] As additional dataset using in both USASpending.gov and FBO, we converted the FAR Product and Service Codes[12] to RDF using LODRefine.[13]

Data resulting from transformation to RDF was interlinked both internally and with external datasets. Internal linking was done in order to fuse equivalent instances of public contracts and business entities (both contracting authorities and bidders). Deduplication was performed using data processing unit for UnifiedViews that wraps Silk link discovery framework.[14] The output links were merged using data fusion component of UnifiedViews.[15] Links to external resources were created either by using code-based URI templates in transformation to RDF or by instance matching based on converted data. The use of codes as strong identifiers enabled automatic generation of links to FAR codes and North American Industry Classification System 2012,[16] two controlled vocabularies used to express objects and kinds of public contracts. Instance matching was applied to discover links to DBpedia and OpenCorporates.[17] Links to DB-

[6] `http://usaspending.gov/`

[7] `https://www.fbo.gov/`

[8] `https://github.com/opendatacz/USASpending2RDF`

[9] `https://github.com/cygri/tarql`

[10] `ftp://ftp.fbo.gov/datagov/`

[11] `https://github.com/opendatacz/FBO2RDF`

[12] `http://www.acquisition.gov/`

[13] `http://code.zemanta.com/sparkica/`

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

[15] Developed previously for ODCleanStore, the predecessor of UnifiedViews [5].

[16] `http://www.census.gov/eos/www/naics/index.html`

[17] `https://opencorporates.com/`

pedia were created for populated places referred to from postal addresses in the U.S. procurement dataset. In this case, the employed Silk linkage rule was based on comparison of normalized ZIP codes and pre-filtering possible matches by transitively-expanded Wikipedia category for populated places in the U.S. OpenCorporates was used as target for linking bidding companies. The task was carried out using batch reconciliation API of OpenCorporates via interface in LODRefine. Links were established based on pre-filtering by jurisdiction and fuzzy matching on normalized legal name, with which company is registered in respective jurisdiction. In all cases of instance matching samples of resulting links were verified by manual scrutiny, in order to estimate linking accuracy.

Eventually, from the whole dataset, those contracts were selected from which the award information was *assumed* to be available between the participant result submission date and the LDMC (Know@LOD 2014) event, but not *before* the result submission date. This part of data became the *testing* dataset, while the contracts for which the award information was known at the time of publishing the data on the LDMC web page became the *training* dataset.

Numbers on the procurement linked dataset, as available to the LDMC participants, is available from `http://knowalod2014.informatik.uni-mannheim.de/en/linked-data-mining-challenge/dataset-description-task-a/`

3.2 Evaluation data

As evaluation data, we understand those data objects in the ordinal prediction task for which the target attribute value could be determined during the evaluation period.

Using automated linkage rule we managed to match public contracts to their award information (including the number of tenders) only in 6 cases out of 788 instances in the testing dataset. Using exact matches on normalized contract identifiers, only 2 links were discovered. Changing the linkage rule to take into account more data besides identifiers, such as contract titles, and tweaking configuration of the rule's comparers yielded a slight improvement to 6 matches.

It is likely that the recall of linking contract notices to their corresponding contract award notices is low. This may be caused by a lack of key identification data or too heterogeneous description of contract notice and its award data. Besides the insufficient linking there may be several reasons why contract award notices for the tested contract notices are unavailable. Publication of a contract award notice can lag behind the typical delay following contract award date. In some cases, contract award notices may not be revealed to the public via USASpending.gov. Some of the contracts in question may have been cancelled prior to award.

Given the unsuccessful attempt in automatic link discovery, we considered to find matching contract award data manually. However, searching for example contracts revealed that a manual approach will not increase the number of found matching contracts.

Ultimatelly, due to the minimal size of the evaluation dataset we decided to abandon the intended evaluation by comparing predicted data to actual data.

4 The Linked Data Mining Challenge results

The 2014 edition of the Linked Data Mining Challenge received only a single submission, which addressed the ordinal prediction task. Unfortunately, as mentioned previously, only 6 objects (contracts) were eventually available for evaluation. Accuracy has thus not even been calculated, since its informativeness would be extremely low.

Instead, the characteristics of the data mining process (decision trees) were briefly discussed during the LDMC session of the workshop. The trees were rather large and their discriminative features specific for individual contract authorities and low-level CPV codes. A high risk of overfitting and lack of interesting insights for experts was thus stated.

As was the case for the previous year's LDMC, it proved surprisingly difficult for the participant to provide the challenge's results in valid CSV. This time the submission was delivered in PDF instead of CSV.

5 Questionnaire survey

In order to wrap up the LDMC effort over the last three years, and to get better insights into the problems and opportunities of such a kind of challenge, a questionnaire survey was carried out as follow-up to the 2014 edition. It was exclusively targeted at the participants of the Know@LOD 2014 workshop,[18] i.e. people with high probability of being aware of the call for the challenge (which was associated with the call for the workshop as such, although with a slightly different timeframe).

5.1 Questionnaire structure

The questions in it partly followed the lifecycle of challenge participation; the respondents were asked about:

1. Attendance to ESWC 2014 in general, and specifically to the Know@LOD 2014 workshop.
2. Whether they noticed the LDMC call early enough to consider taking part in the challenge.
3. (For those who did notice the call:) Whether they downloaded the challenge data and had a look at them, and for which track.
4. (For those who did not download any data:) What prevented them from considering participation. The options were:
   - Lack of expertise in data mining

[18] The list of invitees included all registered Know@LOD 2014 participants, except the LDMC organizers themselves and the author of the only submission (to simplify the questionnaire – it could then explicitly seek the reasons for *not* participating; the author of the submission already provided sufficient feedback in the submitted paper and the talk, so there was no significant information loss).

   - Lack of expertise in linked data
   - The data not interesting enough from the linked data point of view
   - The data was not interesting enough from the data mining point of view
   - The subject domains not topical for them
   - Too short time till the deadline (given other duties)
   - The overhead associated with challenge participation not paying of by the benefit of having a paper at Know@LOD
   - Other reason (to be specified in a comment field)
5. (For those who did download some data:) What prevented them from submitting the results to LDMC. The options were:
   - The overall data structure too complicated to make sense of
   - The data not interesting enough from the linked data point of view
   - The data was not interesting enough from the data mining point of view
   - The participants tried to pre-process the data but ran into technical trouble
   - The data miner did not return any meaningful hypotheses (for already pre-processed data)
   - The work was progressing fine, but the participants were not able to finish it in time
   - Other reason (to be specified in a comment field)
6. Any suggestions to make a future LDMC edition more attractive. The options were:
   - Announcing it much farther ahead the submission deadline
   - Making the data format simpler
   - Including more external links to make it more attractive for linked data researchers
   - Addressing a different community than semantic web researchers
   - Changing the subject domain/s to more appealing one/s.

A general textual field for suggestions or comments was also offered.

5.2 Questionnaire results

Due to the small size of the survey pool and importance of the textual comments, the results are summarized in text rather than in aggregate tables:

- The questionnaire was sent to 22 subjects. We received 6 responses (which means a 27% response rate). Of these, one was however explicitly from a different person than the original addressee, namely, a participant to the *previous* LDMC to whom the e-mail was forwarded by the original addressee. The remaining 5 participated in ESWC 2014, and 4 of them also attended the Know@LOD 2014 workshop. Four people (of which three participated at Know@LOD 2014 and 1 was the previous LDMC participant) declared that they had seen the LDMC 2014 announcement when it appeared.

- Only one respondent downloaded the data. S/he downloaded both datasets (public procurement and researcher collaboration). As for reasons for not participating, s/he complained of ‘data structure too complicated’ and ‘data not interesting from the linked data point of view’. S/he specified in the comment: “I could have tried my framework but the data format was requiring some data preprocessing that would have taken more time. Also, I think the data were not connected enough to the LOD cloud (or at least this was my impression when I had a look at them).”
- *All* the other three respondents aware of the LDMC announcement chose ‘too short time’ as one of the reasons for not participating. Other reasons were diverse:
  - ‘overhead of participation’
  - ‘data not interesting from the DM point of view’
  - ‘other’, specified as (by the previous edition participant): “My main reason for not participating was that the data (public procurement) has not changed much since the previous LDMC (at ECML2013). I participated in that challenge and got the feeling that the problem is very hard and might even be impossible to solve with this amount of data. Also since I did not develop new methods, I did not feel I could do a better job than I did for the first challenge”.
- As for suggestions to make a next edition more attractive, interestingly, all the options were used roughly homogeneously. The 14 hints from the 6 respondents were distributed as follows: 3x ‘announce sooner’, 3x ‘simpler data format’, 3x ‘include more links’, 3x ‘address a different community’ and 2x ‘change the subject domain’. Specific textual hints were:
  - “Communities to address: Natural Language Processing, Machine Learning (for NLP).”
  - “Data Miners, Artificial Intelligence researchers and students”
  - “I would probably generalise the challenge, so that people with different expertise could be interested in applying their own techniques.”
  - “I would focus on attracting machine learning and data mining researchers in the following ways: i) apart form the bit that’s unusual (using RDF/LOD) the task should be a juicy, traditional machine learning task: binary classification, non-skewed data, lots of information in the dataset, lots of instances; ii) the dataset should be a real knowledge graph, with very complex and deep graph structure, which is difficult to translate to traditional features; iii) the task should hint at a solution to a real, unsolved AI problem (like handwriting recognition and image classification); iv) a sample pipeline (perhaps in rapidminer, matlab or scipy) should be provided, to show how to import RDF data, translate it to features and run it through a simple classifier. This pipeline should work as a baseline, but be easy to beat, to inspire the researcher; v) the target score should provide a good long-term moving target. For instance, the best test error on MNIST is 0.0023, and it’s still improving.”
  - “I think the dataset should be larger and should have an even clearer prototypical machine learning problem. I.e. a straightforward binary classi-

> fication problem with a large amount of data. Have the focus on learning from RDF, but have very typical tasks."

It should be mentioned that the content of the questionnaire was already influenced by the discussion that was held on this topic in the end of the Know@LOD workshop itself. This may explain why the choice of prepared answers relatively well covered the suggestions by the respondents (several of them already took part in this discussion) and few completely new proposals appeared within the unstructured part of the questionnaire.

6 Conclusions and suggested next steps

In this paper, we have given an overview on the tasks, preparation of data, submitted results, and, finally, an ex post questionnaire survey of the 2014 Linked Data Mining Challenge.

Although the tangible results of the challenge events proper are rather scanty, by coupling our hands-on experience with its preparation with the questionnaire results we managed to collect useful insights that can help organizers of future challenges related to data mining over linked data. The most important lesson probably is that the activity, as it was conceived so far, aimed to reach too far. As linked data are inherently tough to process in any way (due to their heterogeneity, varying quality and rich structure), it is currently risky to associate them with a data mining problem that is either somewhat non-standard (like the ordinal prediction for the number of tenders) or too vague (such as the exploratory task). It is also risky to rely on data that might not be available, in the core of the result evaluation. Additionally, the role of extensive external data that could be 'easily' linked on demand, is important; this is, however, a problem to be first tackled at the level of linked data as such (before moving on the data mining ground).

The key principles proposed for the organizers of similar challenges (which are likely to arise, given the popularity of both LD and DM) are as follows:[19]

1. Assure that the core underlying data are *well curated*, and as clean as possible. Data mining tools are more demanding in this respect than common data querying applications.
2. Assure that there is a potential of continuously bringing in new data by *linking* the original entities to new datasets. This new data should be of such nature that they potentially could have discriminatory power over the original dataset.
3. While the data cannot escape from being complex, the *data mining problem* should be as straightforward as possible.

[19] Although we already had several of them (especially, the data quality and interlinking, timely availability and the baseline showcase) in mind when preparing the challenge, we acknowledge that we did not manage to fully implement them, for various reasons; the most important one was the rather limited capacity allocated to this task, being only a minor thread within the dissemination workpackage WP10.

4. Dependency on *live data* for *result evaluation* should be reduced. (This partly disqualifies public data, since withholding a secret part of it is impossible.)
5. Give the participants ample *time* to get familiar to the format and structure of the data.
6. Provide a *baseline showcase* for the whole process. This might help attract not only computer scientists but also (for topics of societal importance, such as public procurement) journalists and NGO activists; while their use of DM tools might be basic, they could benefit from understanding the domain and data in depth.
7. Do not underestimate 'superficial marketing' at the level of sticky slogans. The prime participants are students, and they should not fear that the task as whole is boring.

Acknowledgements

The preparation of the Linked Data Mining Challenge and of this paper has been partially supported by the EU ICT FP7 under No. 257943, LOD2 project.

References

1. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
2. Gerben Klaas Dirk de Vries. Graph kernels for task 1 and 2 of the linked data data mining challenge 2013. In *International Workshop on Data Mining on Linked Data (DMoLOD 2013)*, 2013.
3. Dongkyu Jeon and Wooju Kim. Development of prediction model for linked data based on the decision tree. In *3rd Workshop on Knowledge Discovery and Data Mining meets Linked Open Data*, 2014.
4. Eneldo Loza Mencia, Simon Holthausen, Axel Schulz, and Frederik Janssen. Using data mining on linked open data for analyzing e-procurement information - a machine learning approach to the linked data mining challenge 2013. In *International Workshop on Data Mining on Linked Data (DMoLD 2013)*, 2013.
5. Jan Michelfeit and Tomáš Knap. Linked Data Fusion in ODCleanStore. In *International Semantic Web Conference (Posters and Demos)*, 2012.
6. Trina Myers, Jarrod Trevathan, Dianna Madden, and Tristan O'Neil. Linked Data for Cross-disciplinary Collaboration Cohort Discovery. In *First International Workshop on Linked Data for Information Extraction (LD4IE 2013)*, 2013.
7. Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In *International Semantic Web Conference*, 2014. To appear.
8. Vojtěch Svátek, Jindřich; Mynarz, and Petr Berka. Linked Data Mining Challenge (LDMC) 2013 Summary. In *International Workshop on Data Mining on Linked Data (DMoLD 2013)*, 2013.
9. Vojtěch Svátek, Jindřich Mynarz, David Chudán, Jakub Klímek, Łukasz Grzybowski, Mateusz Jarmużek, Krzysztof Wecel, Lorenz Bühmann, and Sander van der Waal. Application of data analytics methods on linked data in the domain of public sector contracts (LOD2 Deliverable D9a.3.1). `http://svn.aksw.org/lod2/D9a.3.1/public.pdf`, 2014.