Profiling the Linked (Open) Data

Blerina Spahiu

Università degli Studi di Milano-Bicocca
spahiu@disco.unimib.it

Abstract. The number of datasets published as Linked (Open) Data is constantly increasing with roughly 1000 datasets as of April 2014. Despite this number of published datasets, their usage is still not exploited as they lack comprehensive and up to date metadeta. The metadata hold significant information not only to understand the data at hand but they also provide useful information to the cleansing and integration phase. Data profiling techniques can help generating metadata and statistics that describe the content of the datasets. However the existing research techniques do no cover a wide range of statistics and many challenges due to the heterogeneity nature of Linked Open Data are still to overcome. This paper presents the doctoral research which tackles the problems related to Linked Open Data Profiling. We present the proposed approach and also report the initial results.

Keywords: Linked Open Data, Profiling, Data Quality, Topical Classification

1 Problem Statement

With 12 datasets in 2007, the Linked Open Data cloud has grown to more than 1000 datasets as of April 2014 [17], a number that is constantly increasing. The datasets to be published need to adopt a series of rules in a way that it would be simple for them to be searched and queried [3]. The datasets should be published adapting W3C standarts in RDF[1] format and made available for SPARQL[2] endpoint queries. Adapting these rules allow different data sources to be connected by typed links which are useful to extract new knowledge as linked datasets do not have the same information. Even though the Linked Open Data is considered a gold mine, its usage is still not exploited as understanding a large and unfamiliar RDF dataset is still a key challenge. As a result of a lack of comprehensive descriptive information the consumption of these dataset is still low. Data profiling techniques support data consumption and data integration with statistics and useful metadata about the content of the datasets. While traditional profiling techniques solve many issues these techniques can not be applied to heterogeneous data such as Linked Open Data. Data profiling techniques in the context of Linked Open Data are very important for different tasks:

[1] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
[2] http://www.w3.org/TR/rdf-sparql-query/

Complex schema discovery. Schema complexity leads to difficulties to understand and access databases. Schema summaries provide users a concise overview of the entire schema despite its complexity.

Ontology / schema integration. Ontologies published on the Web, even for datasets in similar domains can have differences. Data profiling techniques can help understanding the overlap between ontologies and help in the process of ontology creation, maintenance and integration.

Big knowledge bases and provide a landscape view. Data profiling techniques can help identifying some core knowledge patterns (KP) which reveal a piece of knowledge in a domain of interest.

Inspect large datasets to find quality issues. Data profiling tools allow the inspection of large datasets for detecting quality issues, by identifying the cases that do not follow business rules, outliers detection, residuals, etc.

Data integration. To perform a data integration process, one should consider schema mapping, the process of discovering relationships between schemas. Profiling techniques can reveal mappings between classes and properties, helping the integration process.

Entity summarization. Finding features that best represent the topic/s of a given dataset can help not only the topical classification of the dataset but also understanding the semantic of the information found in the data.

Data visualization for summarization. Profiling techniques can support data visualization tools to visualize large multidimensional datasets by displaying only a small and concise summary of the most relevant and important features, enhancing the comprehension of the user by allowing him to dig into the data by zooming in or out the provided summary.

In this proposal we will focus on the profiling techniques to summarize the content of a dataset and reveal data quality problems. Moreover we will propose profiling techniques combined with data mining algorithms to find useful and relevant features to summarize the content of datasets published as Linked Open Data and also techniques that reveal quality issues in the data. The dataset summarization can be used not only to detect if the dataset is useful or not, but also to provide useful information to the cleansing and integration phase.

2 Related Works

Statistics and summaries can help to describe and understand large RDF data. Most of the existing profiling tools, support traditional databases which are homogeneous and have a well-defined schema. These techniques can not be applied to Linked Open Data due to their heterogeneity and the lack of a well-defined schema. As it will be discussed most of the existing techniques to profile Linked Open Data are limited in few statistics and summaries covering only one task.

Roomba [1] is a framework to automatically validate and generate descriptive dataset profiles. The extracted metadata are grouped into four categories (general, access, ownership or provenance) depending on the information they hold. After metadata extraction some validation and enrichments steps are performed.

Metadata validation process identifies missing information and automatically corrects them when it is possible. As an outcome of the validation process, a report is produced which can be automatically sent to the dataset maintainer.

The ExpLOD [8] tool is used to summaries a dataset based on a mechanism that combines text labels and bisimulation contractions. It considers four RDF usages that describe interactions between data and metadata, such as class and predicate instantiation, class and predicate usage on which it creates RDF graphs. It also uses the `owl:sameAs` links to calculate statistics about the interlinking between datasets. The ExpLOD summaries are extracted using SPARQL queries or algorithms such as partition refinement.

RDFStats [9] generates statistics for datasets behind SPARQL endpoint and RDF documents. It is built on Jena Semantic Framework and can be executed as a stand-alone process, important to optimize SPARQL queries. These statistics include the number of anonymous subjects and different types of histograms; URIHistogram for URI subject and histograms for each property and the associated range(s). It uses also methods to fetch the total number of instances for a given class, or a set of classes and methods to obtain the UIRs of instances.

LODStats [2] is a profiling tool which can be used to obtain 32 different statistical criteria for datasets from Data Hub. These statistics describe the dataset and its schema and include statistics about number of triples, triples with blank nodes, labeled subjects, number of `owl:sameAs` links, class and property usage, class hierarchy depth, cardinalities etc. These statistics are then represented using Vocabulary of Interlinked Datasets (VoID)[3] and Data Cube Vocabulary[4].

ProLOD [5] is a web based tool which analyzes the object values of RDF triples and generates statistics upon them such as data type and patterns distribution. In ProLOD the type detection is performed using regular expression rules and normalized patterns are used to visualize huge numbers of different patterns. ProLOD also generates statistics on literal values and external links. ProLOD++[5] which is an extension of ProLOD is also a browser based tool which implements several algorithms with the aim to compute different profiling, mining or cleansing tasks. In the profiling task are included processes to find frequencies and distribution of distinct subjects, predicates and objects, range of the predicates etc. ProLOD++ can also identify predicates combinations that contain only unique values as key candidates to distinctly identify entities. The implementation of mining tasks cover processes such as synonym and inverse predicate discovering, association rules on subjects, predicates and objects, etc. It also performs some cleansing tasks such as auto completions of new facts for a given dataset, ontology alignment in identifying predicates which are synonym or identifying cases where the pattern usage is over specified or underspecified.

Profiling as the activity of providing insights through the data, is not only about providing statistics about value distribution, null values etc, but also is referred to the process of finding and extracting information patterns in the data.

[3] http://www.w3.org/TR/void/
[4] http://www.w3.org/TR/vocab-data-cube/
[5] https://www.hpi.uni-potsdam.de/naumann/sites/prolod++/app.html

In the area of schema summarization Knowledge Patterns (KP) can be defined as a template to organise meaningful knowledge [6]. The approach in [15] identifies an abstraction named dataset knowledge architecture that highlights how a dataset is organized and which are the core knowledge patterns (KP) we can retrieve from that dataset. These KPs summarise the key features of one or more datasets, revealing a piece of knowledge in a certain domain of interest.

Encyclopedic Knowledge Patterns (EKP) [12] are some knowledge patterns introduced to extract core knowledge for entities of a certain type from Wikipedia page links. EKPs are extracted from the most representative classes describing a concept and containing abstraction of properties. The use of EKPs that supports exploratory search is showen in Aemoo[6] to enrich query results with relevant knowledge coming from different data sources in the Web [13].

In order to understand complex datasets, [4] introduces Statistical Knowledge Pattern (SKP) to summarize key information about an ontology class considering synonymity between two properties of a given class. An SKP is stored as an OWL ontology and contains information about axioms derived or not expressed in a reference ontology but can be promoted applying some statistical measures.

As shown, the actual profiling tools provide schema based statistics like the class/property usage, incoming/outgoing links etc, but none of the existing works is focused in providing summarization of the content of the dataset and also apply techniques to profile its quality. Author in [7] propose an approach to profile the Web of Data, but in difference from this, the proposed approach profiles Linked Data in terms of its quality and summarize datasets in terms of its topic.

3 Research Plan

The contribution of this PhD in the area of Linked Open Data Profiling covers (i) generating new statistics that are not covered by the state of the art techniques (ii) new algorithms to overcome the challenges to perform profiling in the LOD, and (iii) the development of a methodology on how to perform profiling tasks. In the following we will give an overview of the methodology which we want to follow in order to accomplish the contribution we want to make in the field.

New statistics for Linked Data Profiling

While much effort is done as described in the state of the art, the generated statistics are limited in some basic statistics such as the number of triples, number of classes/ properties that are used in a dataset, the datatypes or `sameAs` links used, etc. Datasets hold much more interesting information which might be hidden, but at the same time, this information could be useful for the consumer of the dataset. As data profiling is referred to the activity of providing useful descriptive information, new techniques on how to extract the hidden information should be developed. Our intent is to develop automatic approaches to generate new statistics and knowledge patterns to provide dataset summary and inspect its quality. Different data mining techniques, such as association rule mining, can be used to discover and extract patterns and dependencies in the dataset. These

[6] http://wit.istc.cnr.it/aemoo/

patterns might provide useful information especially to detect errors and inconsistencies in spatial data (*consistency* quality dimension). Implementation of different approaches for outlier detection, like distance/deviation/depth-based, evolutionary techniques, etc. could provide insight about abnormalities in the underlying data. Other techniques such as clustering, classification, aggregation, dimensionality reduction or spatial data summarization might help to provide concise and accurate dataset summarization and inspect quality dimensions mentioned in [16]. We intend to further investigate the topical classification of LInked Open Data. The datasets published as LOD cover a wide range of topics but they lack metadata that describe the topical category, so the users have difficulties deciding if the dataset is relevant for their interest or not. For each of the dataset published as LOD a label for the topical category was manually assigned [17]. The datasets have only one label for the topical category while often two or more topics are needed to describe a dataset. The actual topical classification of datasets in the LOD is limited to eight categories, while a more fine-grained topical classification might provide more useful information.

Overcoming Profiling Challenges

As another contribution in this research we want to tackel the profiling challenges described in [11]. Traditional profiling task can not be applied to Linked Data due to their heterogeneity. Heterogeneity can appear in different forms such as different formats or query languages called syntactic heterogenity. Linked Open Data can be represented in different formats, stored in different storage architectures also the data encoding schemes may vary. This is referred to as schematic heterogeneity. Datasets published as LOD might use different vocabularies, to describe synonymous terms. [11] referred semantic heterogeneity as the discovery of semantic overlap of the data. Traditional data profiling tools can not be used to profile Linked Open Data as they suppose data to be homogeneous stored in a single repository, while Linked Open Data are neither homogeneous nor stored in a single repository. Also as the number of the datasets published is increasing the need to adapt and optimise profiling techniques to support huge amount of data is also high. A good approach when dealing with large datasets, is to improve the profiling performance running the calculation of statistics and patterns extraction in parallel. We also plan to adapt some data mining techniques to deal with high dimensionality data, such as Linked Open Data.

Methodology to Profile Linked Open Data

As another contribution of this research we intend to develop a methodology on how to perform profiling tasks. This methodology would classify profiling tasks depending on the purpose and also provide guidelines to appropriate select the tasks needed by the user.

4 Preliminary Results

This PhD work is now at the second year. As a first step we measured the value of Linked Open Data, profiling the data published as Open Data from the Italian Public Administrations. In this work we profiled the adoption of Linked Open

Data best practices and local laws by the Italian Public Administration calculating a compliance index considering three quality dimensions for the published data; *completeness, accuraccy* and *timeliness* [18].

As mentioned in the Sec. 3, the main contribution of this research is to provide new techniques for dataset summarization and new statistics about the data. ABSTAT[7] is a framework which can be used to summarise linked datasets and at the same time to provide statistics about them. The summary consists of Abstract Knowledge Patterns (AKPs) of the form <*subjectType, predicate, objectType*> which represent the occurrence of triples <sub,pred,obj> in the data, such that subjectType is a minimal type of sub and objectType is a minimal type of obj. The ABSTAT summaries can help users comparing in which of two datasets a concept is described with richer and diverse properties, and also help detecting errors in the data such as missing or datatype diversity, etc [14]. ABSTAT can also be used to fix the domain and range information for properties. Either the domain or the range is unspecified for 585 properties in DBpedia Ontology and AKPs can help us in determining at least one domain and one range for the unspecified properties. For example, for the property `http://dbpedia.org/ontology/governmentType` in DBpedia we do not have information about the domain. With our approach we can derive 7 different AKPs meaning that we can derive 7 domains for this property.

We further investigated one of the challenges still present in the Linked Open Data datasets, topic classification. We built the first automatic approach to classify LOD datasets into the topical categories that are used by the LOD cloud diagram. For the classification we considered eight feature sets; vocabulary, classes and properties usage, local class/property names, text from `rdfs:lable`, top-level domain and in and out degree. In Table 1, are shown the results training three classifiers k-NN, Naive Bayes and Decision Tree on three balancing approaches, no sampling, down and up sampling and two normalization techniques considering the binary occurrence and the relative term occurence for each term or vocabulary. Our approach achieves an accuracy of 81,62% [10].

Table 1. Results of combined feature sets. Best three results in bold.

Classification Approach	Accuracy in %			
	ALL_{bin}	ALL_{rto}	$NoLab_{bin}$	$NoLab_{rto}$
k-NN (no sampling)	74.93	71.73	76.93	72.63
k-NN (down sampling)	52.76	46.85	65.14	52.05
k-NN (up sampling)	74.23	67.03	71.03	68.13
J48 (no sampling)	**80.02**	77.92	79.32	79.01
J48 (down sampling)	63.24	63.74	65.34	65.43
J48 (up sampling)	79.12	78.12	79.23	78.12
Naive Bayes (no sampling)	21.37	71.03	**80.32**	77.22
Naive Bayes (down sampling)	50.99	57.84	70.33	68.13
Naive Bayes (up sampling)	21.98	71.03	**81.62**	77.62

A deep literature study for the tools which are used to profile LOD has been taken. We analyzed existing tools in terms of the goal they are used for,

[7] http://abstat.disco.unimib.it/

techniques, input, output, approach, automatization information, license etc, with the aim to have a complete view of the existing approaches and techniques for profiling which helps us in determining new statistics or new techniques. This deepen study will also help us for the third contribution classifying profiling tasks and creating a general methodology for each task depending on the use case.

5 Lessons Learned, Open Issues and Future Work

The main contribution of this PhD work is to address the challenges mentioned in Sec. 3 to built a framework for profiling the Linked Open Data in order to give insights of the data, despite their heterogeneous nature. To evaluate the validity of the proposed approach or the results achieved is very difficult as in the filed of LOD profiling there is no Gold Standard, thus is very difficult to compare with others. For this issue, we want to further explore how these new statistics or summarization allow to improve the performance of the actual profiling techniques and tools, e.g. how profiling tasks can improve full-text search etc. To evaluate the validity of the proposed profiling techniques to summarise datasets, as pattern discovery is not trivial, humans will evaluate the validity of the summarization in terms of relatedness and informativeness. We intend to provide to users a list of statistics and ask them which in their opinion is more important to support profiling of Linked Open Data. The evaluation of the performance of profiling tasks is very difficult, which still remains an open issue on which I am currently working.

The ABSTAT framework provides some contributions in summarising Linked Open Data, and detecting quality issues. We are working to enrich this framework with other statistics and to apply it to unstructured data such as microdata.

Regarding the topical classification of LOD datasets, we will consider the problem for multi-label classification. As the datasets in the LOD cloud are unbalanced a two stage approach might help, while a classifiers chain which makes a prediction for one class after the other could address the multi-lable problem. Up till now in our experiments we have not exploited RDF links beyond datasets in and out degree, so link-based classification techniques could be applied to further investigate the content of a dataset.

Acknowledgements

This research has been supported in part by FP7/2013-2015 COMSODE (under contract number FP7-ICT-611358). I would like to thank my supervisor Assoc. Prof Andrea Maurino, my supervisor during my visiting period Prof. Dr Christian Bizer, Asst. Prof Matteo Palmonari, Dr. Anisa Rula for their priceless suggestions and also the anonymous reviewers for their helpful comments.

References

[1] A. Assaf, R. Troncy, and A. Senart. Roomba: An extensible framework to validate and build dataset profiles. In *The 2nd International Workshop on Dataset PRO-*

FIling and fEderated Search for Linked Data (PROFILES '15) co-located with ESWC 2015, Portorož, Slovenia, May 31 - June 1, 2015., pages 32–46, 2015.

[2] S. Auer, J. Demter, M. Martin, and J. Lehmann. Lodstats - an extensible framework for high-performance dataset analytics. In *18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012.*

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[4] E. Blomqvist, Z. Zhang, A. L. Gentile, I. Augenstein, and F. Ciravegna. Statistical knowledge patterns for characterising linked data. In *Proceedings of the 4th Workshop on Ontology and Semantic Web Patterns co-located with ISWC 2013, Sydney, Australia, October 21, 2013.*

[5] C. Böhm, F. Naumann, Z. Abedjan, D. Fenz, T. Grütze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with prolod. In *Workshops Proceedings of the 26th ICDE 2010, March 1-6, 2010, Long Beach, California, USA.*

[6] A. Gangemi and V. Presutti. Towards a pattern science for the semantic web. *Semantic Web*, 1(1-2):61–68, 2010.

[7] A. Jentzsch. Profiling the web of data. *Proceedings of the 8th Ph. D. retreat of the HPI research school on service-oriented systems engineering*, page 101, 2014.

[8] S. Khatchadourian and M. P. Consens. Explod: Summary-based exploration of interlinking and RDF usage in the linked open data cloud. In *ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010*, pages 272–287, 2010.

[9] A. Langegger and W. Wöß. Rdfstats - an extensible RDF statistics generator and library. In *Database and Expert Systems Applications, DEXA, International Workshops, Linz, Austria, August 31-September 4, 2009*, pages 79–83, 2009.

[10] R. Meusel, B. Spahiu, C. Bizer, and H. Paulheim. Towards automatic topical classification of lod datasets. In *Proceedings of the 24th International Conference on World Wide Web, LDOW Workshop, 2015, Florence,Italy, May 18-22, 2015.*

[11] F. Naumann. Data profiling revisited. *SIGMOD Record*, 42(4):40–49, 2013.

[12] A. G. Nuzzolese, A. Gangemi, V. Presutti, and P. Ciancarini. Encyclopedic knowledge patterns from wikipedia links. In *The Semantic Web - ISWC 2011 Bonn, Germany, October 23-27, 2011, Proceedings, Part I*, pages 520–536, 2011.

[13] A. G. Nuzzolese, V. Presutti, A. Gangemi, A. Musetti, and P. Ciancarini. Aemoo: exploring knowledge on the web. In *Web Science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2-4, 2013*, pages 272–275, 2013.

[14] M. Plamonari, A. Rula, R. Porrini, A. Maurino, B. Spahiu, and V. Ferme. Abstat: Linked data summaries with abstraction and statistics. In *European Semantic Web Conferenze 2015 (ESWC2015) Portoroz, Slovenia, 31th May - 4th June 2015.*

[15] V. Presutti, L. Aroyo, A. Adamou, B. A. C. Schopman, A. Gangemi, and G. Schreiber. Extracting core knowledge from linked data. In *Proceedings of the COLD 2011, Bonn, Germany, October 23, 2011*, 2011.

[16] A. Rula and A. Zaveri. Methodology for assessment of linked data quality. In *Proceedings of the 1st Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany, September 2nd, 2014.*, 2014.

[17] M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *The Semantic Web - ISWC 2014, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 245–260, 2014.

[18] G. Viscusi, B. Spahiu, A. Maurino, and C. Batini. Compliance with open government data policies: An empirical assessment of italian local public administrations. *Information Polity*, 19(3-4):263–275, 2014.