Guest Editorial Preface

Special Issue on Dataset Profiling and Federated Search for Linked Data

Elena Demidova, WAIS Group, ECS, University of Southampton, Southampton, UK
Stefan Dietze, L3S Research Center, Hannover, Germany
John G. Breslin, Insight Centre for Data Analytics, NUI Galway, Galway, Ireland
Julian Szymański, Gdansk University of Technology, Gdańsk, Poland

INTRODUCTION

The Web of Data and in particular Linked Open Data (LOD) has seen tremendous growth recently. In just three years, the number of publicly available datasets within the LOD cloud increased by more than three times from 295 datasets in 2011 to over 1000 datasets in 2014 (Schmachtenberg et al., 2014). These datasets cover a wide variety of domains from publications and life sciences to government and user-generated content. Furthermore, significant amounts of data are currently emerging through Web markup languages such as RDFa, Microdata and Microformats. Nevertheless, despite this growth, the actual take up, interlinking and reuse of Web of Data remains limited, and is often focused on well-known reference datasets, such as DBpedia (Bizer et al., 2013), YAGO (Suchanek et al., 2007) and Freebase (Bollacker et al., 2008).

One of the main obstacles preventing broad reuse of LOD datasets in end user applications is the lack of efficient and scalable methods for formulating and distributing keyword queries across the Web of Data. This problem is further accelerated by a lack of trust in the quality of search results retrieved using federated search over distributed third-party datasets, particularly in the absence of provenance information. As demonstrated by a number of recent studies, especially on Linked Data automatically generated from semi-structured and unstructured sources, there are a number of quality problems, including extraction and interlinking errors, as well as incompleteness of schema information, as demonstrated for example by (Acosta et al., 2013) and (Paulheim & Bizer, 2014). Hence, dataset profiling, quality analytics and selection represent inherent challenges for query distribution.

In particular, dataset selection for a given task is currently hindered by the lack of trustworthy and current information about the nature, characteristics, currentness and suitability of particular datasets. As the LOD cloud includes data from a variety of domains spread across more than 1000 datasets containing billions of entities and facts and is constantly evolving, manual assessment of dataset features is neither feasible nor sustainable, often leading to sparse and outdated dataset metadata. This is, for example, apparent with DataHub, the largest dataset registry for open datasets in general and LOD in particular (Buil-Aranda et al., 2013). Hence, given the dynamic and evolving nature of the LOD cloud, particular focus should be paid to the development of scalable automated approaches that can facilitate frequent assessment and profiling of large scale datasets to enable efficient selection of suitable datasets for query federation and search.

PAPERS INCLUDED IN THE SPECIAL ISSUE

For this special issue, we accepted five papers in the area of dataset profiling and selection with the authors coming from: Croatia (3), Greece (5), Belgium (6), France (3), Ireland (4) and Chile (1) integrating two territories: Europe and South America. This geographical distribution reflects an increasing research interest in the topic of Linked Data profiling and reuse around the globe. From a topical perspective, this special issue contains a selection of articles addressing a number of important issues in the space of dataset content profiling and dataset selection. From the profiling perspective, we include articles performing text analysis and dataset catalogue creation. Furthermore, several important aspects of dataset selection are addressed, including their evolution, connectivity, scalable discovery and quality. In the following, we provide a more detailed overview of the included papers.

Dataset Content Profiling: Text Analysis and Keyword Networks

To examine the types of words found in the texts of data collections, which of course can assist with dataset content profiling and therefore selection, Beliga et al. describe a collection-oriented keyword extraction method called "Selectivity-Based Keyword Extraction" (SBKE) that extracts keywords from texts and represents them as a graph. A node selectivity value is calculated and used to rank keyword candidates. The authors show that their SBKE measure outperforms typical node centrality measures such as in-degree/out-degree and betweenness. One of the advantages of such an extraction is the fact that it is based on the statistical/structural information of the graph, and does not require any linguistic knowledge.

Dataset Selection: Evolution and Connectivity

Mountantonakis et al. also look at the profiles of datasets, but more so in terms of their evolution over time and long-term connectivity. In their paper, they describe the requirements and challenges associated with building and maintaining a semantic warehouse, in particular in relation to its quality and value. The authors propose a set of metrics for quantifying the connectivity of such a warehouse, which is in turn related to how well it can answer advanced queries via its endpoints. An extension of the VoID (Vocabulary of Interlinked Datasets) vocabulary is used to provide improved endpoint profiles, which can be leveraged by federated search applications. The evolution of a semantic warehouse after subsequent reconstructions can also be analysed using the metrics to assist with quality checking and looking at how the warehouse has evolved.

Dataset Selection: Scalable Discovery

The challenge of how to select and query data from multiple datasets is tackled by Vander Sande et al. in their paper on hypermedia-based algorithms for finding and selecting datasets via Linked Data interfaces. As mentioned in the previous paper, federated search requires accurate information on the available data sources, and this paper looks at a data source discovery approach based on hypermedia links and controls. Quantitative metrics are used to evaluate the hypermedia-based approach, with interfaces to discover multiple large real-world datasets shown to discover each other within 7 minutes and with a 50% reduction in execution time.

Dataset Selection: Quality

Access to multiple datasets is not enough, as one needs to determine how good the quality of the data in those datasets actually is, so as to assist with choosing the best data sources from all those available. Assaf et al. propose an objective assessment framework for assessing Linked Data quality. The authors propose a set of objective quality indicators that are use-case agnostic and can be checked automatically. A thorough survey of Linked Data quality tools is also presented, which shows varying levels of quality indicators that are available from these tools. A more extensive (and extendable) Linked Data quality tool is presented that can assist dataset owners to determine the quality of their

own datasets, with some pointers towards how to improve these datasets. Dataset seekers can also benefit from this framework as it allows them to select the desired datasets from a ranked set. The tool is used to evaluate the quality of the overall Linking Open Data cloud, highlighting some common issues with many datasets as evidenced by their various quality scores.

Dataset Content Profiling: Catalogue Creation

A means for cataloguing the content of public SPARQL endpoints is enabled through certain queries facilitated by SPARQL 1.1, and can be aligned to VoID descriptions, although not all endpoints support this functionality as of yet. In the final paper of this special issue, we come full circle to the issue of dataset content profiling, where Hasnain et al. describe how public endpoints can self-describe their content, assisting in the creation of dataset catalogues. These catalogues could then be discovered and queried by agents looking for particular content. The authors highlight the fact that many endpoints do not yet support SPARQL 1.1, while others may only return partial responses or none at all for rejected/expensive queries. Two goals of the paper are to investigate how well public endpoints can perform content self-descriptions, using VoID profiles as comparison, and to create a best-effort dataset catalogue based on available results from public endpoints.

FUTURE RESEARCH DIRECTIONS

As discussed above and as documented by the papers in this special issue, it is apparent that a range of different techniques and methods can contribute towards solving the problem of dataset profiling, assessment, federated search and query distribution. This involves scalable and efficient means for analysing dataset characteristics of all kinds, vocabularies for the representation of dataset metadata in an interoperable and reusable way, and finally, the large scale exploitation of such metadata to guide query distribution and federated search on the Web of Data. So far, considerable work has been conducted to address specific facets of the aforementioned aspects. However, this work is often carried out in constrained or domain-specific settings.

While existing works represent very important steps towards the generation of precise metadata and provenance information and thus, hopefully, broader reuse of the Web of Data, we still observe significant potential for further research in this area. This research is especially important when it comes to the Web-scale exploitation of such efforts. A wide range of existing works investigates the assessment of specific dataset characteristics, such as their dynamics, schema compliance and topic coverage (e.g. (Ellefi et al., 2016) and (Fetahu et al., 2014)). On the top of these effort, shared vocabularies beyond general frameworks such as DCAT or VoID, for sharing the broad variety of dataset characteristics on the Web, are still missing. In addition, scalability and efficiency seem to represent crucial issues, specifically given the evolving nature of the Web of Data, requiring the recurring and periodic computation of dataset metadata to ensure its currentness.

The lack of a common means for representation and scalability issues with current approaches seems to be among the reasons why dataset metadata and profiles are not yet widespread. In turn, federated search or dataset selection can only be informed very partially and with limited degrees of success by such forms of dataset metadata. This involves both automated approaches for dataset selection and search as well as semi-automatic approaches guided by recommendation methods. Finally, even though most works in this area in general, and in this special issue in particular, seem somewhat naturally confined to Linked (Open) Data on the Web, they mostly leave aside new emerging forms of Web data. This includes significant amounts of data emerging through Web markup languages such as RDFa, Microdata and Microformats, which open up new, Web-scale forms of structured data on the Web.

Overall, the areas discussed above represent challenging and promising research directions in the field of LOD and Web of Data in the coming years, and have significant potential to support LOD and Web of Data reuse and spawn a wide variety of novel real-world applications.

ACKNOWLEDGMENT

The work on this special issue was supported by the KEYSTONE COST Action IC1302. Elena Demidova's work was partially funded by the H2020-MSCA-ITN-2014 WDAqua (grant agreement 64279) and ALEXANDRIA (ERC 339233). John G. Breslin's work has been funded by Science Foundation Ireland under grant number SFI/12/RC/2289 (INSIGHT).

REFERENCES

Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., & Lehmann, J. (2013). Crowdsourcing Linked Data Quality Assessment.*Proceedings of the 12th International Semantic Web Conference*.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S. (2009). DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2009(7), 154–165.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge.*Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD '08)* (pp. 1247-1250). ACM, New York, NY, USA. doi:10.1145/1376616.1376746

Buil-Aranda, C., Hogan, A., Umbrich, J., & Vandenbussch, P.-Y. (2013). SPARQL Web-Querying Infrastructure: Ready for Action?*Proceedings of the 12th International Semantic Web Conference*.

Ellefi, M. B., Bellahsene, Z., Dietze, S., & Todorov, K. (2016, May).Intension-based Dataset Recommendation for Data Linking. *Proceedings of the 13th Extended Semantic Web Conference (ESWC 2016)*, Heraklion, Crete.

Fetahu, B., Dietze, S., Nunes, B. P., Casanova, M. A., & Nejdl, W. (2014). A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. *Proceedings of the 11th Extended Semantic Web Conference, LNCS* (Vol. 8465, pp. 519-534). Springer International Publishing. doi:10.1007/978-3-319-07443-6_35

Paulheim, H., & Bizer, C. (2014). Improving the Quality of Linked Data Using Statistical Distributions. *International Journal on Semantic Web and Information Systems*, *10*(2), 63–86. doi:10.4018/ijswis.2014040104

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the Linked Data Best Practices in Different Topical Domains. *Proceedings of ISWC 2014*, *LNCS* (Vol. 8796, pp. 245-260). doi:10.1007/978-3-319-11964-9_16

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007, May 8-12). Yago: a core of semantic knowledge.*Proc. of the 16th International Conference on World Wide Web, WWW 2007*, Banff, Alberta, Canada (pp. 697– 706).