Identification of Similar Entities in the Web of Data

Michal HOLUB*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
michal.holub@stuba.sk

The Semantic Web presumes structured, machine-readable data published in widely accepted open standards on the Web. Once this becomes true, new automated applications will emerge that will automate many of our tasks. So, the key to this lies in high quality semantic data.

Already, such data is being created and linked together, thus forming the Linked Data cloud or the Web of Data. Currently, there are few hundreds of such datasets covering wide range of domains (public domain, music, movies, biology, bibliography, etc.). However, so far only few attempts were made to actually pursue the idea of wider adoption of such datasets in various web-based applications (e.g. search, recommendation, and navigation).

In our research we focus on discovering links between various entities in the Linked Data cloud. Our method can also be used for automatic concept map construction which has applications in the knowledge management domain. For this purpose we use unstructured data from the Web, which we transform to concepts and discover links between them. We proposed such solution in the domain of programming languages and related technologies [6]. Resulting concept map is usable for recording the technical knowledge and skills of software engineers.

We examine the utilization of such concept maps and Linked Data utilization in order to 1) improve the navigation in a digital library based on what the user has already visited [8], 2) find similarities between scientists and authors of research papers and recommend them to the readers browsing a digital library [2], 3) analyze Linked Data graphs and find identical entities [5], 4) enhance the displayed articles based on linking entities to DBpedia [4] and recommend additional interesting information to the reader within a digital library, 5) enable users to search using queries in English language [3].

Linked Data are being used in various datasets forming a Linked Data Cloud. In the center of this cloud there are two main datasets: DBpedia [1] and YAGO [7]. Both

* Supervisor: Mária Bieliková, Institute of Informatics, Information Systems and Software Engineering

Spring 2016 PeWe Workshop, April 2, 2016, pp. 55–56.

use Wikipedia as their primary source of information. The goal of these datasets is to extract and define as many entities as possible, so that others can link to them.

For the purpose of describing the knowledge of software developers we propose the creation of a concept map. The map is composed of a set of concepts representing various technologies and principles the developers are familiar with.

Building an adaptive web-based application using a domain model based on linked data enables us to utilize the relationships to recommend related entities (e.g. in the domain of learning materials), or to help the user navigate in a large information space (e.g. in large digital libraries containing millions of authors, papers and conferences which may overwhelm the user). We can also use the relationships to help the user in the search process. Since the Linked Data cloud has the form of a large graph we are able to answer complex queries, which are difficult to solve using traditional keyword-based approach.

We evaluate the models and methods of their creation directly by comparing them to existing ones or by evaluating facts from them using domain experts. Moreover, we evaluate the models indirectly by incorporating them in adaptive personalized web-based systems and measure the improvement in the experience of users (i.e. they get better recommendations, search results, etc.).

References

[1] Auer S., Lehmann J.: What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In: *The Semantic Web: Research and Applications*, LNCS Vol. 4519. Springer, (2007), pp. 503-517.

[2] Chen H., Gou L., Zhang X., Giles C.L.: CollabSeer: A Search Engine for Collaboration Discovery. In: *Proc. of the 11th Annual Int. ACM/IEEE Joint Conf. on Digital Libraries*, ACM Press, (2011), pp. 231-240.

[3] Chong W., Xiong M., Zhou Q., Yu Y.: PANTO: A Portable Natural Language Interface to Ontologies. In: *The Semantic Web: Research and Applications*, LNCS Vol. 4519. Springer (2007), pp. 473-487.

[4] Exner P., Nugues P.: Entity Extraction: From Unstructured Text to DBpedia RDF Triples. In: *The Web of Linked Entities Workshop*, Boston, USA, (2012).

[5] Halpin H., Hayes P.J., McCusker J.P., McGuinness D.L., Thompson H.S.: When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data. In: *Proc. of the 9th Int. Semantic Web Conf. on The Semantic Web – Volume Part I*. Springer (2010), pp. 305-320.

[6] Holub, M., Kuric, E., Bieliková, M.: Modeling the Knowledge of Developers using the Lightweight Semantics. In: *ZNALOSTI 2012: Sborník příspěvků 11 ročníku konference, Matfyzpress,* (2012), pp. 21–30. In Slovak

[7] Suchanek F.M., Kasneci G., Weikum G. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: *Proc. of the 16th int. Conf. on World Wide Web*, ACM Press, (2007), pp. 697-706.

[8] Waitelonis J., Harald S.: Augmenting Video Search with Linked Open Data. In: *Proc. of Int. Conf. on Semantic Systems*, Verlag der TU Graz, Austria, (2009).