Discovering Relationships between Entities in Web-based Digital Libraries

Michal HOLUB*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
holub@fiit.stuba.sk

A lot of information can be found on the Web; however, it is mostly intended for humans. If the computers could better understand the data it would enable us to make more intelligent web applications which could extract new facts, better adapt to users' needs and make searching easier and faster. This requires representing the data together with semantics and relationships, which is the focus of our research.

In particular, we focus on the domain of web-based digital libraries which contain a lot of information useful for other researchers. The main entities with which we deal are authors, papers, conferences and publications. The problem is that data is scattered across many web portals so we need to integrate it. This requires finding relationships between the entities from different sources.

Apart from obvious relationships, e.g. person writes a paper, there are few explicitly expressed relationships between these entities. Relationships bring another dimension to the data which we can use for filtering, finding similarities or differences, etc. Search engines working with the Web of Data with semantics can provide more precise results for the queries, especially when asking questions about entities.

Relationships between entities and objects are also essential for their integration and creating mashups of things. We can use it with exploratory search when we create an overview of the target domain from web objects. There are various types of relationships which we can find between entities. The most interesting relationships are created based on the interaction of users with web objects. These relationships may not have a meaningful name but can express relatedness of the objects.

In our research we focus on building a platform enabling users doing research more efficiently. The platform is based on the semantic data and collaboration by its users.

The base is a domain model of digital libraries represented by RDF triples. This allows us to record relationships and their types. We crawl various web portals (e.g. ACM Digital Library, Springer, DBLP, CiteULike) and parse information about

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

Spring 2012 PeWe Workshop, April 1, 2012, pp. 59–60.

entities, which we subsequently integrate into one dataset. Apart from metadata about scientific articles this dataset also contains user-generated content like tags.

Next, we transform the data according to the Linked Data principles and we discover relationships between various entities. We plan to use a reasoner to derive new facts.

Contribution of our approach is deriving relationships between entities also according to the behavior of users when doing research (e.g. when searching for a paper on selected topic). We transform the actions he makes into implicit feedback and connect the entities accordingly. These relationships can vary per user so we get more configurations of the domain model which can be used for different groups of users.

The process of relationship discovery introduces new research challenges such as verification and validation of relationships, their weighting and ranking. Newly discovered relationships enrich the domain model which later leads to improvement of search, personalization and recommendation processes.

We have done an experiment involving automatic generation of facets for filtering the entities in digital libraries represented in our domain model. Nowadays, faceted interfaces are being successfully used to browse textual data [1]. This is useful when we have a long list of search results which we want to narrow down. In this experiment we have not yet included the relationships.

We applied the information retrieval methods on the attributes of the entities (e.g. title of an article, name of the conference). After tokenization, stemming and stop-words removal we counted the document frequencies for each term in the collection (here, the document means the entity). As facets we used the terms occurring in the interval 10 % - 50 % of the entities.

The evaluation showed that this method is appropriate for attributes which do not have a wide range of values, e.g. the names of the events (workshops and conferences) where the results are satisfactory. However, it is not applicable to attributes with values from a very wide range, e.g. article titles. In this case, the precision of our method was only 25 % (we used the domain expert who chose the meaningful facets which we then compared to the results obtained using our method).

In the future work we would like to incorporate the relationships into the process of facet generation, which was partially done in [2]. Then, we will observe the behavior of users using the faceted search and derive new relationships which will improve the domain model.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.B31:B52.

References

[1] Dakka, W., Ipeirotis, P.G., Wood, K.R.: Automatic Construction of Multifaceted Browsing Interfaces. In: *Proc. of the 14th ACM Int. Conf. on Inf. and Knowledge Management (CIKM '05)*, ACM Press New York, NY, USA, pp. 768–775.

[2] Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A Browser for Heterogeneous Semantic Web Repositories. In *The Semantic Web (ISWC 2006)*, LNCS 4273, Springer, pp. 272–285.