

Annotating Web for Users

Jan Dědek, Alan Eckhardt, and Peter Vojtáš

Department of Software Engineering, Faculty of Mathematics and Physics
Charles University in Prague, Czech Republic
{dedek,eckhardt,vojtas}@ksi.mff.cuni.cz

Abstract. Annotating Web for Users

Keywords: Semantic Annotation, User Preferences, Machine Learning

Semantic annotation (SA) is considered to be one of the most important elements in the evolution of the Semantic Web. It can provide great help in the process of data and information integration and it could also be a basis for intelligent search and navigation. Users can easily get all available information to given topic and see or directly navigate to relevant related facts which can be spread over the whole (Semantic) Web.

Web information extraction is often able to extract valuable data. We would like to couple the extraction process with the consecutive semantic annotation (human trained; later operating automatically). Our idea is to split the extraction process in two parts - domain independent and domain dependent.

Our first approach for domain independent intermediate information extraction and semantic annotation is to use the structural similarity in web pages containing large number of table cells and for each cell a link to detailed pages. This is often presented in web shops and on pages that presents more than one object (product offer). Each object is presented in a similar way and this fact can be exploited [2].

Our second approach for Web information extraction is targeted to textual pages. It is based on deep linguistic analysis produced by Czech linguistics and NLP tools. We use a machine learning method based on Inductive Logic Programming. The learning procedure is capable construct linguistic extraction rules, which can be easily visualized, understood and adapted by human [1].

Although automated and machine learning based methods for Information Extraction and Semantic Annotation can significantly reduce the amount of necessary human work, it is clear that it will be never possible avoid it completely. Therefore it is useful to heave a fancy tool that would be attractive to human users providing them an easy way of doing manual semantic annotation [4]. Thus these users could benefit from having the content of their interest semantically annotated with as low cost as possible. To reduce the costs, which means to reduce the amount of manual annotations per user, we plan to develop a community site with a semantic repository [5] where annotation work can be shared amongst users and assisted by automated tools.

User preference learning is also loosely connected to Semantic web. It is not the core aspect, but without knowing the users needs, the semantic web wouldn't be able to improve user experience. We are investigating the area of preference learning. Since this area consists of many diverse topics, we currently concentrate on content based learning from user's ratings, but the focus shifts towards interpreting user behaviour. Our main motivation is preference learning. The main focus is to learn the preference of one specific user - we are not interested in "the average user".

We are working with the assumption of having annotated data in the database or other repository. Semantic web initiative enables the annotation of data. Annotated data allows us to use the attributes of the objects that may be difficult to find. The attributes may be stored on different web sites; the pages may have complicated structure, etc.

The paradigm of content based learning is to understand what properties an object should have in order to be preferred. Knowing which attribute values are desirable is important for many reasons. Being able to visualise the object and to emphasise those desired properties is the user interface part [4] Another advantage is the use of preferences in the database aspect, top-k query answering and indexing techniques [3].

The most recent area is user behaviour mining. This area is much more difficult than using user ratings to assess his/her preferences, because behaviour contains the preferences only implicitly and we cannot be sure whether they are contained at all. However, our preliminary experiments in [6] have shown that it is possible even on real world web shops. Much more investigation is needed in this area for better acquaintance with the subject.

References

1. Dědek, J.: Towards semantic annotation supported by dependency linguistics and ILP. In: (ISWC2010), Part II. LNCS, vol. 6497, pp. 297–304. Springer, Shanghai / China (2010)
2. Eckhardt, A., Horváth, T., Marušćák, D., Novotný, R., Vojtáš, P.: Uncertainty Issues and Algorithms in Automating Process Connecting Web and User, LNCS, vol. 5327. Springer (2008)
3. Fagin, R., Lotem, A., Naor, M.: Optimal aggregation algorithms for middleware. In: 12th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 102–113. PODS '01, ACM, New York, NY, USA (2001)
4. Fišer, D.: Semantic annotation of domain dependent data. Master's thesis, Charles University in Prague, Czech Republic (2011)
5. Lašek, I., Vojtáš, P.: Semantic information filtering - beyond collaborative filtering. In: Proceedings of the 4rd International Semantic Search Workshop. SEMSEARCH '11, ACM, New York, NY, USA (2011)
6. Peska, L., Eckhardt, A., Vojtas, P.: UPComp - a PHP component for recommendation based on user behaviour. In: Submitted to Web Intelligence and Intelligent Agent Technology (WI-IAT) (2011)