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
{alfred.hofmann,ursula.barth,ingrid.haas,frank.holzwarth,
anna.kramer,leonie.kunz,christine.reiss,nicole.sator,
erika.siebert-cole,peter.strasser,lncs}@springer.com
http://www.ksi.mff.cuni.cz/

Abstract. Annotating Web for Users

Keywords: Semantic Annotation, User Preferences, Machine Learning

1 Semantic Annotation

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.

1.1 Web Information Extraction

Web information extraction [1] is often able to extract valuable data. We would like to couple the extraction process with the consecutive semantic annotation (initially human trained; later it should operate 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. [Maruk]

Our second approach for Web information extraction is targeted to textual pages. It is based on deep linguistic analysis (DLA) produced by Czech linguistics and NLP tools. We use a chain of linguistic analyzers prom the PDT and TectoMT projects that processes the text presented on a web page and produces linguistic (syntactic) trees corresponding with particular sentences. These trees serve as a basis of our semantic extraction. We use a machine learning method based on Inductive Logic Programming (ILP). The combination of DLA and

ILP have several benefits: no need of manual selection of learning features, the learning procedure is capable to select relevant parts itself and to construct linguistic extraction rules, which can be easily visualized, understood and adapted by human. [Ddek]

1.2 Collaborative Approach

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 [Fisher]. 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 [Laek] where annotation work can be shared amongst users and assisted by automated tools.

2 User preferences on the web

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 [Vaclav] (see the picture). Another advantage is the use of preferences in the database aspect, top-k query answering and indexing techniques [Fagin].

Our approach is divided into two steps - criteria for each attribute are learnt during the first step, the preference degrees of attributes are aggregated into the overall preference of the whole object during the second step. The first step transforms attributes from "price" to "cheap" and "weight" to "light" (or "heavy", which depends on the user). The second step then computes the overall preference of the object, based on the satisfaction of the (learnt) criteria.

There are various problems addressed. The first kind of problem comes from the preference nature - the sample rated by the user on the web is often very small (up to 50 objects) and the rating is given on a small set of rating, usually 1,2,3,4,5. Another kind of problem arises from the need to have a transparent user preference model, which will allow it to be used in other components of the system - indexing, query processing.

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 [Peska] 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.

Acknowledgments

This work was partially supported by Czech projects: GACR P202/10/0761, GACR-201/09/H057, GAUK 31009 and MSM-0021620838.

References

1. Liu, B.: Web Data Mining. Springer-Verlag (2007), http://dx.doi.org/10.1007/978-3-540-37882-2