Information Extraction and Integration from Heterogeneous Semi-structured Web Sources in the Domain of Used Cars

Radoslaw Oldakowski

Today's Web offers access to a great amount of information about various products and services. However, with the increasing number of different information sources new problems arise. The main challenge to the customer is to find, integrate, and process all the relevant information needed for making a purchase decision. The importance of this challenge rises with the increasing product involvement. In the case of high involvement goods (being of great value/importance to the customer, e.g. cars) people engage in a more extensive decision-making process based on a detailed search for information and comparison of alternatives.

Nowadays, consumers browsing the Web for product information simply bookmark all the relevant pages and then integrate their content manually. Due to the limitations of the human mind, regarding fast processing of large amounts of information, the decision making process, based on manually integrated information, is time-consuming and error-prone. Therefore, this process is mostly limited to a small number of alternatives or to the comparison of just a few features of a given product type.

We address these problems by proposing an architecture for product information aggregation and purchase decision support based on Semantic Web technologies. In our approach, users, while browsing the Web, store product data instead of bookmarks. Having all the information stored in a machine-understandable format allows enhanced information discovery and information sharing among users, as well as automation of sophisticated tasks like detailed matching of consumer preferences with product characteristics based on information from different sources. In our proposal of the Product Information Aggregation and Purchase Decision Support Architecture we restrict our analysis to a certain kind of products from the automotive domain, namely, passenger cars. There are several rationales speaking in favour of this product category which will be further explained during the talk together with various aspects of the data extraction and integration process.

Searching for information objects in the integrated product descriptions is a nontrivial task. For complex objects having multiple properties a perfect match is rarely found. Therefore, the user is also interested in a ranking of objects with respect to specified preferences. Although the SPARQL query language provides structured access to semantically rich data, a flexible framework on a higher abstraction level on top of SPARQL is needed in order to retrieve property values, to calculate their similarity and subsequently to aggregate them into an overall similarity score. Moreover, such a framework should provide means for personalized queries, be able to utilize the knowledge of concept relationships from an underlying ontology as well as offer various similarity computation and aggregation techniques. We meet those requirements by introducing SemMF, which is an easy-to-use, flexible framework for calculating semantic similarity between objects represented as arbitrary RDF graphs.