Obtaining general concepts that represent a set of objects using ontologies

Ferran Mata Arcas

Advisor: Aïda Valls Mateu

2012

Master of Science Thesis
Master on Computer Security and Intelligent Systems

Universitat Rovira i Virgili

iTAKA
Intelligent Technologies for Advanced Knowledge Acquisition

Summary

The amount of information that a user has to deal with nowadays is huge, usually impossible to manage. Because of that that there is a need to reduce the amount of information returned to the user in his searches on the Web. Recommender systems take into account the user's preferences to filter the results and show only a subset of them, the ones relevant for the user. Filtering can be done on the bases of generating a partition of the set of alternatives into clusters. Hence, studying how to generate a general representation of the semantic concepts that represent a certain cluster is needed.

This work is part of the DAMASK (Data Mining Algorithms with Semantic Knowledge) project, and is focused on the last tasks of it. The DAMASK project proposes the use of semantic domain knowledge, represented in the form of ontologies in different tools that are needed to develop Recommender Systems on the Web. The project is conformed in the strategic area of Tourism with the realization of a Web application for the personalised recommendation of touristic destinations. This Master Thesis has been developed within this context. It is based on some previous works done in this research project, mainly focused on the information extraction of relevant data from a domain of structured, semi-structured (e.g. Wikipedia) and unstructured Web resources.

The Master Thesis includes several steps of the DAMASK project. First, the construction of a data matrix that comprehends the cities and their description using a set of heterogeneous attributes. Those attributes have values of multiple types such as semantic, numerical and categorical. This data matrix is built using the data extracted using the tools previously developed in the DAMASK project. Second, there is the need to create and algorithm to compute the similarity between two cities, because one of the main purposes of the DAMASK system is the creation of partitions (clusters) of similar cities, and to do that is necessary a measure of distance. Third, the clustering method chosen to make the partition of cities is the K-means algorithm. At each step of the K-means there is the need to have an average value. In this step, this Master Thesis has developed a new method to generate a multi-valued centroid to represent the semantic attributes of a given cluster. The process of obtaining of general concepts that represent a set of objects is crucial to make these centroids. Using ontologies, we can select the most appropriate and representative concepts for each semantic attribute.

All those techniques are integrated in a Web Recommender System prototype as part of the DAMASK project. The system is designed for non-experienced users with no knowledge on the topic of this work. This recommender system uses the profile of the user that is searching for touristic destinations to show the most appropriate cluster or partition of cities for the user. In other words, the partitions achieved previously are the results that the system can recommend. The user has also the possibility of filter the results if the recommended cluster is too large.

Finally, the results of the clustering system and centroid construction are evaluated separately and in combination with the recommendations that the Web system retrieves to the user. As the reader of this work will see, the results obtained are quite satisfactory.