Data Mining for the Web

Myra Spiliopoulou

Institut für Wirtschaftsinformatik, Humboldt-Universität zu Berlin
myra@wiwi.hu-berlin.de, http://www.wiwi.hu-berlin.de/~myra

Abstract. The web is being increasingly used as a borderless marketplace for the purchase and exchange of goods, the most prominent among them being information. To remain competitive, companies must recognize the interests and demands of their users and adjust their products, services and sites accordingly.
In this tutorial, we discuss data mining for the web. The relevant research can be categorized into (i) web usage mining, focussing in the analysis of the navigational activities of web users, (ii) web text mining, concentrating on the information being acquired by the users and (iii) user modelling, which exploits all information and knowledge available about the web users to build user profiles.

In this tutorial, we discuss advances of data mining for the analysis of data related to the web and its usage. The tutorial is organized similarly to the classical procedere for knowledge discovery.

Problem Formulation

We briefly discuss the different application areas calling for web mining. They range from marketing, on line shopping and information dissemination by public institutions to hypermedia teachware. We abstract the potential strategic aims in each domain into mining goals as: prediction of the users' behaviour within the site, comparison between expected and actual web site usage, adjustment of the web site to the interests of its users. We also review measures of evaluating whether those goals are achieved

Data Preparation

The information sources available to mine the web encompasses web usage logs, web page descriptions and textual contents, web site topology, user registries and questionnaires. We explain shortly how each of the above sources contributes to fullfilling each of our three tasks and then discuss advances on data preparation for web data mining. Web usage data are noisy and of rather poor quality. Particular emphasis is given to techniques that try to improve their quality by heuristics and statistic methods.

We close this entity by discussing the notion of web data warehousing and the exploitation of OLAP technology to assess information on web site usage.

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 588–589, 1999.

Web Data Mining

We first discuss advances on web usage mining, coming from the domains of sequence mining, discovery of association rules and clustering. The clustering paradigm is mostly used when the focus is on the contents of the pages being accessed together. Hence, clustering for web usage mining serves as a bridge to switch to the discussion on web text mining. We shortly present advances on analyzing text in general. Then, the focus shifts on the particularities of web text and on mining web pages and hypertext. Finally, we turn into methods for user profiling, addressing mostly clustering techniques. This subsection* closes with open issues related to the combination of results from all three domains.

Evaluation of the Results

Here, we discuss statistical methods for evaluating the results of sequence miners, text miners and clusterers. These methods are mostly of general purpose. There are also specific evaluation strategies for web mining results, e.g. from the marketing domain.

We then point out issues that impede the evaluation of web mining results. These include the idiosyncracy of web-related data, which forces some researchers to use synthetic data for testing, as well as biases occuring in all collections of user registration data. Open research issues related to the alleviation of those problems are presented next.

Exploitation of the Results

The exploitation of the results in any data mining domain is claimed to be domain specific, thus falling beyond the well-modelled subjects of knowledge discovery. We describe potential ways of exploiting web mining results, namely by reorganizing web sites statically or dynamically and by building personalized services on the basis of user profiles.

We then elaborate on the need for guidelines to transform the results of a data miner into suggestions or hints for web site reorganization or profile-based size construction. We discuss reports addressing the above issues for particular applications.

Evolution

We thus far have spoken about the web and about web mining as for a static world. However, (almost) nothing is less static than the web. Here, we address two issues: (i) the impact of site changes on the results of web mining and (ii) the observation of changes in web usage, caused e.g. by the increasing experience of users or by shifts of interest.

The first issue relates to rules' maintenance. The second issue is relevant to time series analysis and to temporal mining. However, research in those two domains is not oriented towards web usage analysis. More focussed research is needed to deal with the issue of time in the rapidly changing world of the web.