Chapter XIX

The Scent of a Newsgroup: Providing Personalized Access to Usenet Sites through Web Mining

Giuseppe Manco
Italian National Research Council, Italy

Riccardo Ortale
University of Calabria, Italy

Andrea Tagarelli
University of Calabria, Italy

ABSTRACT

Personalization is aimed at adapting content delivery to users’ profiles: namely, their expectations, preferences and requirements. This chapter surveys some well-known Web mining techniques that can be profitably exploited in order to address the problem of providing personalized access to the contents of Usenet communities. We provide a rationale for the inadequacy of current Usenet services, given the actual scenario in which an increasing number of users with heterogeneous interests look for information scattered over different communities. We discuss how the knowledge extracted from Usenet sites (from the content, the structure and the usability viewpoints) can be suitably adapted to the specific needs and expectations of each user.

INTRODUCTION

The term knowledge discovery in databases is usually devoted to the (iterative and interactive) process of extracting valuable patterns from massive volumes of data by
exploiting data mining algorithms. In general, data mining algorithms find hidden structures, tendencies, associations and correlations among data, and mark significant information. An example of data mining application is the detection of behavioural models on the Web. Typically, when users interact with a Web service (available from a Web server), they provide enough information on their requirements: what they ask for, which experience they gain in using the service, how they interact with the service itself. Thus, the possibility of tracking users’ browsing behaviour offers new perspectives of interaction between service providers and end-users. Such a scenario is one of the several perspectives offered by Web mining techniques, which consist of applying data mining algorithms to discovery patterns from Web data. A classification of Web mining techniques can be devised into three main categories:

- **Structure mining.** It is intended here to infer information from the topology of the link structure among Web pages (Dhyani et al., 2002). This kind of information is useful for a number of purposes: categorization of Websites, gaining an insight into the similarity relations among Websites, and developing suitable metrics for the evaluation of the relevance of Web pages.

- **Content mining.** The main aim is to extract useful information from the content of Web resources (Kosala & Blockeel, 2000). Content mining techniques can be applied to heterogeneous data sources (such as HTML/XML documents, digital libraries, or responses to database queries), and are related to traditional Information Retrieval techniques (Baeza-Yates & Ribeiro-Neto, 1999). However, the application of such techniques to Web resources allows the definition of new challenging application domains (Chakrabarti, 2002): Web query systems, which exploit information about the structure of Web documents to handle complex search queries; intelligent search agents, which work on behalf of users based both on a description of their profile and a specific domain knowledge for suitably mining the results that search engines provide in response to user queries.

- **Usage mining.** The focus here is the application of data mining techniques to discover usage patterns from Web data (Srivastava et al., 2000) in order to understand and better serve the needs of Web-based applications and end-users. Web access logs are the main data source for any Web usage mining activity: data mining algorithms can be applied to such logs in order to infer information describing the usage of Web resources. Web usage mining is the basis of a variety of applications (Cooley, 2000; Eirinaki & Vazirgiannis, 2003), such as statistics for the activity of a Website, business decisions, reorganization of link and/or content structure of a Website, usability studies, traffic analysis and security.

Web-based information systems depict a typical application domain for the above Web mining techniques, since they allow the user to choose contents of interest and browse through such contents. As the number of potential users progressively increases, a large heterogeneity in interests and in the knowledge of the domain under investigation is exhibited. Therefore, a Web-based information system must tailor itself to different user requirements, as well as to different technological constraints, with the ultimate aim of personalizing and improving users’ experience in accessing the system. Usenet turns out to be a challenging example of a Web-based information system, as it encompasses a very large community, including government agencies, large universities, high schools, and businesses of all sizes. Here, newsgroups on new topics are
Related Content

Web Information Resources Vis-à-Vis Traditional Information Services
www.igi-global.com/chapter/web-information-resources-vis-vis/21992?camid=4v1a

A New Approach in Cloud Computing User Authentication
www.igi-global.com/chapter/a-new-approach-in-cloud-computing-user-authentication/140906?camid=4v1a

Estimation of a Priori Decision Threshold for Collocations Extraction: An Empirical Study
www.igi-global.com/article/estimation-of-a-priori-decision-threshold-for-collocations-extraction/100051?camid=4v1a
A New Framework for Intelligent Semantic Web Services Based on GAIVAs
www.igi-global.com/chapter/new-framework-intelligent-semantic-web/40420?camid=4v1a