IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB8939

Chapter XX

Critical and Future Trends in Data Mining: A Review of Key Data Mining Technologies/ Applications

Jeffrey Hsu
Fairleigh Dickinson University, USA

ABSTRACT

Every day, enormous amounts of information are generated from all sectors, whether it be business, education, the scientific community, the World Wide Web (WWW), or one of many readily available off-line and online data sources. From all of this, which represents a sizable repository of data and information, it is possible to generate worthwhile and usable knowledge. As a result, the field of Data Mining (DM) and knowledge discovery in databases (KDD) has grown in leaps and bounds and has shown great potential for the future (Han & Kamber, 2001). The purpose of this chapter is to survey many of the critical and future trends in the field of DM, with a focus on those which are thought to have the most promise and applicability to future DM applications.

MAJOR TRENDS IN TECHNOLOGIES AND METHODS: WEB MINING

Web mining is one of the most promising areas in DM, because the Internet and WWW are dynamic sources of information. Web mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the WWW (Etzioni, 1996). The main tasks that comprise Web mining include retrieving Web documents, selection and processing of Web information, pattern discovery in sites and across sites, and analysis of the patterns found (Garofalis, Rastogi, Seshadri & Shim, 1999; Kosala & Blockeel, 2000; Han, Zaiane, Chee, & Chiang, 2000).

Web mining can be categorized into three separate areas: web-content mining, Web-structure mining, and Web-usage mining. Web-content mining is the process of extracting knowledge from the content of documents or their descriptions. This includes the mining of Web text documents, which is a form of resource discovery based on the indexing of concepts, sometimes using agent-based technology. Web-structure mining is the process of inferring knowledge from the links and organizations in the WWW. Finally, Web-usage mining, also known as Web-log mining, is the process of extracting interesting patterns in Web-access logs and other Web-usage information (Borges & Levene, 1999; Kosala & Blockeel, 2000; Madria, Bhowmick, Ng, & Lim, 1999).

Web-content mining is concerned with the discovery of new information and knowledge from web-based data, documents, and pages. According to Kosala and Blockeel (2000), there are two main approaches to Web-content mining: an information retrieval view and a database (DB) view. The information retrieval view is designed to work with both unstructured (free text, such as news stories) or semistructured documents (with both HTML and hyperlinked data), and attempts to identify patterns and models based on an analysis of the documents, using such techniques as clustering, classification, finding text patterns, and extraction rules (Billsus & Pazzani, 1999; Frank, Paynter, Witten, Gutwin & Nevill-Manning, 1998; Nahm & Mooney, 2000). The other main approach, which is to content mine semi-structured documents, uses many of the same techniques as used for unstructured documents, but with the added complexity and challenge of analyzing documents containing a variety of media elements (Crimmins & Smeator, 1999; Shavlik & Elassi-Rad, 1998).

There are also applications that focus on the design of languages, which provide better querying of DBs containing web-based data. Researchers have developed many *web-oriented query languages* that attempt to extend standard DB query languages such as SQL to collect data from the WWW, e.g., WebLog and WebSQL. The TSIMMIS system (Chawathe et al., 1994) extracts data from heterogeneous and semistructured information sources and correlates them to generate an integrated DB representation of the extracted information (Maarek & Ben Shaul, 1996; Han, 1996; Meldelzon, Mihaila, & Milo, 1996; Merialdo, Atzeni, & Mecca, 1997).

Other applications focus on the building and management of *multilevel or multilayered DBs*. This suggests a multilevel-DB approach to organizing web-based information. The main idea behind this method is that the lowest level of the DB contains primitive semistructured information stored in various Web repositories, such as hypertext documents. At the higher level(s), metadata or generalizations are extracted from lower levels and organized in structured collections such as relational or object-oriented DBs. Kholsa, Kuhn, and Soparkar (1996) and King and Novak (1996) have done research in this area.

Web-structure mining. Instead of looking at the text and data on the pages themselves, Web-structure mining has as its goal the mining of knowledge from the structure of websites. More specifically, it attempts to examine the structures that exist between documents on a website, such as hyperlinks and other linkages. For instance,