Web Mining Overview

Bamshad Mobasher
DePaul University, USA

INTRODUCTION

In the span of a decade, the World Wide Web has been transformed from a tool for information sharing among researchers into an indispensable part of everyday activities. This transformation has been characterized by an explosion of heterogeneous data and information available electronically, as well as increasingly complex applications driving a variety of systems for content management, e-commerce, e-learning, collaboration, and other Web services. This tremendous growth, in turn, has necessitated the development of more intelligent tools for end users as well as information providers in order to more effectively extract relevant information or to discover actionable knowledge.

From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. Web mining (i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage) is the collection of technologies to fulfill this potential.

In this article, we will summarize briefly each of the three primary areas of Web mining—Web usage mining, Web content mining, and Web structure mining—and discuss some of the primary applications in each area.

BACKGROUND

Knowledge discovery on and from the Web has been characterized by four different but related types of activities (Kosala & Blockeel, 2000):

1. **Resource Discovery:** Locating unfamiliar documents and services on the Web.
2. **Information Extraction:** Extracting automatically specific information from newly discovered Web resources.
3. **Generalization:** Uncovering general patterns at individual Web sites or across multiple sites.
4. **Personalization:** Presentation of the information requested by an end user of the Web.

The goal of Web mining is to discover global as well as local structures, models, patterns, or relations within and between Web pages. The research and practice in Web mining has evolved over the years from a process-centric view, which defined Web mining as a sequence of tasks (Etzioni, 1996), to a data-centric view, which defined Web mining in terms of the types of Web data that were being used in the mining process (Cooley et al., 1997).

MAIN THRUST

The evolution of Web mining as a discipline has been characterized by a number of efforts to define and expand its underlying components and processes (Cooley et al., 1997; Kosla & Blockeel, 2000; Madria et al., 1999; Srivastava et al., 2002). These efforts have led to three commonly distinguished areas of Web mining: Web usage mining, Web content mining, and Web structure mining.

Web Content Mining

Web content mining is the process of extracting useful information from the content of Web documents. Content data correspond to the collection of facts a Web page was designed to convey to the users. Web content mining can take advantage of the semi-structured nature of Web page text. The HTML tags or XML markup within Web pages bear information that concerns not only layout but also the logical structure and semantic content of documents.

Text mining and its application to Web content have been widely researched (Berry, 2003; Chakrabarti, 2000). Some of the research issues addressed in text mining are topic discovery, extracting association patterns, clustering of Web documents, and classification of Web pages. Research activities in this field generally involve using techniques from other disciplines, such as information retrieval (IR), information extraction (IE), and natural language processing (NLP).

Web content mining can be used to detect co-occurrences of terms in texts (Chang et al., 2000). For example, co-occurrences of terms in newswire articles may show that gold frequently is mentioned together with copper when articles concern Canada, but together with silver when articles concern the US. Trends over time also may be discovered, indicating a surge or decline in interest in certain topics, such as programming languages like Java. Another application area is event detection, the identifi-

cation of stories in continuous news streams that correspond to new or previously unidentified events.

A growing application of Web content mining is the automatic extraction of semantic relations and structures from the Web. This application is closely related to information extraction and ontology learning. Efforts in this area have included the use of hierarchical clustering algorithms on terms in order to create concept hierarchies (Clerkin et al., 2001), the use of formal concept analysis and association rule mining to learn generalized conceptual relations (Maedche & Staab, 2000; Stumme et al., 2000), and the automatic extraction of structured data records from semi-structured HTML pages (Liu, Chin & Ng, 2003). Often, the primary goal of such algorithms is to create a set of formally defined domain ontologies that represent precisely the Web site content and to allow for further reasoning. Common representation approaches are vector-space model (Loh et al., 2000), descriptive logics (i.e., DAML+OIL) (Horrocks, 2002), first order logic (Craven et al., 2000), relational models (Dai & Mobasher, 2002), and probabilistic relational models (Getoor et al., 2001).

Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes and hyperlinks as edges connecting between two related pages. Web structure mining can be regarded as the process of discovering structure information from the Web. This type of mining can be divided further into two kinds, based on the kind of structural data used (Srivastava et al., 2002); namely, hyperlinks or document structure. There has been a significant body of work on hyperlink analysis, of which Desikan et al. (2002) provide an up-to-date survey. The content within a Web page also can be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents (Moh et al., 2000) or on using the document structure to extract data records or semantic relations and concepts (Liu, Chin & Ng, 2003; Liu, Grossman & Zhai, 2003).

By far, the most prominent and widely accepted application of Web structure mining has been in Web information retrieval. For example, the Hyperlink Induced Topic Search (HITS) algorithm (Klienberg, 1998) analyzes hyperlink topology of the Web in order to discover authoritative information sources for a broad search topic. This information is found in authority pages, which are defined in relation to hubs as their counterparts: Hubs are Web pages that link to many related authorities; authorities are those pages that are linked by many good hubs. The hub and authority scores computed for each Web page indicate the extent to which the Web page serves as a hub pointing to good authority pages or as an authority on a topic pointed to by good hubs.

The search engine Google also owes its success to the PageRank algorithm, which is predicated on the assumption that the relevance of a page increases with the number of hyperlinks pointing to it from other pages and, in particular, of other relevant pages (Brin & Page, 1998). The key idea is that a page has a high rank, if it is pointed to by many highly ranked pages. So, the rank of a page depends upon the ranks of the pages pointing to it. This process is performed iteratively until the rank of all the pages is determined.

The hyperlink structure of the Web also has been used to automatically identify Web communities (Flake et al., 2000; Gibson et al., 1998). A Web community can be described as a collection of Web pages, such that each member node has more hyperlinks (in either direction) within the community than outside of the community.

An excellent overview of techniques, issues, and applications related to Web mining, in general, and to Web structure mining, in particular, is provided in Chakrabarti (2003).

Web Usage Mining

Web usage mining (Cooley et al., 1999; Srivastava et al., 2000) refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal of Web usage mining is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests.

The primary data sources used in Web usage mining are log files automatically generated by Web and application servers. Additional data sources that also are essential for both data preparation and pattern discovery include the site files and meta-data, operational databases, application templates, and domain knowledge.

The overall Web usage mining process can be divided into three interdependent tasks: data preprocessing, pattern discovery, and pattern analysis or application. In the preprocessing stage, the clickstream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. In the pattern discovery stage, statistical, database, and machine learning operations are performed to obtain possibly hidden patterns reflecting the typical behavior of users, as well as summary statistics on Web resources, sessions, and users. In the final stage of