Trends in Web Content and Structure Mining

Anita Lee-Post
University of Kentucky, USA

Haihao Jin
University of Kentucky, USA

INTRODUCTION

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. This area of research is fast-developing today, drawing attention and interests from both researchers and practitioners. The tremendous growth of information available on the Web and the recent interest in e-commerce have accounted for this phenomenon (Kosala & Blockeel, 2000).

BACKGROUND

Depending on the nature of the data to be mined, Web mining can be categorized into three areas: Web content, Web structure, and Web usage mining (Srivastava, Cooley, Deshpande, & Tan, 2000).

- Web content mining is the discovery or retrieval of useful information from the content of the Web including text, images, audio, video, and other forms of content that make up the Web pages such as symbolic, metadata, and hyperlink data. Text and hypertext content are the most common sources of data for content mining. The information extracted holds the key to search engine operations. The resulting information mined is represented as an index. A key word supplied by a user is matched against this index to retrieve relevant information for the user. An ideal index is one that links every string, word, phrase, tune and image on the Web to all the pages that contain them (Linoff & Berry, 2001).
- Web structure mining is the discovery of useful information from the underlying hyperlink structures of the Web. The structure is represented as a graph showing how pages or documents within a site and between sites are linked (Broder, Kumar, Maghoul, Raghavan, Rajagopalan & Stata, 2000). An ideal graph is one that maps all links connecting every document on the entire Web (Linoff & Berry, 2001). By analyzing the topology of the Web, information such as the popularity and richness of a document can be revealed. Links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or the variety of topics covered in the document. Such information enhances the usefulness of a search engine, adding popularity and richness to the relevancy of information retrieved.
- Web usage mining is the discovery of useful information from users' usage patterns. Usage data in the form of pages visited, duration of the visit, navigation paths, browse/click pattern, etc. is available from Web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks and scrolls, and any other data generated by the interaction of users and the Web. Usage data is represented as user profiles. An ideal user profile is generated from continually updating records of an individual user's interactions with the Web including, among other things, sites visited, paths taken, queries issued, documents read, and items purchased. The user profile is particularly useful for e-commerce companies to track and predict customer behavior on their Web sites.

We will discuss in detail past research contributions with respect to Web content and structure mining next. Research efforts relating to Web usage mining will be covered in a separate chapter.

MAIN THRUST

Web Content Mining

Web content mining focuses on the discovery or retrieval of useful information from Web content/data/documents. The contributions of Web content mining can be evaluated on two fronts: information retrieval or search result mining and information extraction or Web page content mining (Pal, Talwar, & Mitra, 2002).

INFORMATION RETRIEVAL

Search result mining is about information retrieval on the Web, a task performed by search engines. The goal of information retrieval is to find what you are looking for precisely. The usefulness of a search engine hinges on its ability to retrieve a relevant subset of documents expediently from a large collection of Web pages based on a user query.

The most commonly used technique for search result mining is Web document classification or text categorization. Classification is the process of assigning an item to a class with a certain degree of confidence. In search result mining, classification amounts to assigning keywords to Web documents with varying degrees of confidence. A Web document classified in this way makes possible its subsequent retrieval based on keyword-based searches. According to Kosala and Blockeel (2000), text categorization can be represented in various forms, including the frequencies of specific words, phrases, concept categories, and named entities. The relevance of the page retrieved is measured by its rank: the most relevant pages appear at the top of the list of returned pages. Various rules are used by different search engines in ranking Web pages. For instance, some search for the frequency and location of keywords or phrases in the Web page document, while others scan the META tag, title field, headers and text near the top of the document.

An alternative to the ranked list approach for information retrieval is document clustering. Rather than presenting users with a ranked list, document clustering partitions the retrieval results in sets or clusters based on the topical homogeneity of Web documents retrieved. The topical homogeneity is expressed as a similarity metric constructed by analyzing textual contents, hyperlinks, and/or co-citation patterns of Web documents (He, Zha, Ding, & Simon, 2002).

Recent developments in search result mining include image classification, multimedia information retrieval and cross-language information retrieval. Yanai (2003) used a large number of images from the Web as training data in generic image classification. Meghini, Sebastiani and Straccla (2001) combined similarity and semantic based methods to retrieve text and images from forms, content, and structure of the Web. Kwork (2000) and Chen, Chau and Yeh (2004) proposed means for cross-language text retrieval and multilingual text mining.

Information Extraction

Web page content mining is about information extraction, originally the task of locating specific information from a natural language document. The goal of information extraction is to extract relevant facts from documents as opposed to retrieving relevant documents to satisfy a user's information need as in information retrieval. With information extraction, information is pulled from texts of heterogeneous formats, including PDF files, emails, and Web pages, and organized into a single homogeneous form such as tables within a relational database. Essentially the Web content is converted into a database that end-users can search or organize into taxonomies, allowing them to wade through and cope with the overwhelming amount of digital information by breaking up the Web into small, more manageable pieces (Adams, 2001).

The most commonly used technique for Web page content mining is natural language processing which recognizes words, attribute values and even concepts in a restricted domain (Adams, 2001). The process of pulling relevant information from Web documents and restructuring it as a database is known as feature extraction. For example, pages about automboile engine design can be scanned to extract features such as engine capacity and fuel consuming rate to forecast trends in these areas. Feature extraction on the Web has been used by e-commerce sites to compare prices for similar products, a service known as comparative shopping. Until more advanced technology that extracts information from unstructured text is available, the current approach to comparative shopping is restricted to analyzing structured contents in the form of XML tags on the Web (Linoff & Berry, 2001).

Recent developments in Web page content mining include Web table mining, opinions or reputation extraction, information fusion, and concept mining. Yang and Luk (2002) proposed means to extract information from tables embedded inside Web pages. Dave, Lawrence & Pennock (2003) extracted opinions of a product from the Web for product reviews. Morinaga, Yamanishi, Tateishi and Fukushima (2002) used Web content mining techniques to extract information about a target product's reputation from the Web. Etzioni, Cafarella and Downey (2004) extracted and fused information from multiple documents in a domain independent and scalable manner. Liu, Chin and Ng (2003) and Loh, Wives, Leandro and Oliveira (2000) worked with topic-specific concepts and definitions to discover concept-based knowledge in text extracted from the Web.

Web Structure Mining

Web structure mining uses the hyperlink structure of the Web to yield useful information, including definitive pages specification, hyperlinked communities identification, Web pages categorization, and Web site completeness evaluation.