Web Mining for Public E–Services Personalization

W

Penelope Markellou
University of Patras, Greece

Angeliki Panayiotaki
University of Patras, Greece

Athanasios Tsakalidis
University of Patras, Greece

INTRODUCTION

Over the last decade, we have witnessed an explosive growth in the information available on the Web. Today, Web browsers provide easy access to myriad sources of text and multimedia data. Search engines index more than a billion pages and finding the desired information is not an easy task. This profusion of resources has prompted the need for developing automatic mining techniques on Web, thereby giving rise to the term "*Web mining*" (Pal, Talwar, & Mitra, 2002).

Web mining is the application of data mining techniques on the Web for discovering useful patterns and can be divided into three basic categories: *Web content mining*, *Web structure mining,* and *Web usage mining*. Web content mining includes techniques for assisting users in locating Web documents (i.e., pages) that meet certain criteria, while Web structure mining relates to discovering information based on the Web site structure data (the data depicting the Web site map). Web usage mining focuses on analyzing Web access logs and other sources of information regarding user interactions within the Web site in order to capture, understand and model their behavioral patterns and profiles and thereby improve their experience with the Web site.

As citizens requirements and needs change continuously, traditional information searching, and fulfillment of various tasks result to the loss of valuable time spent in identifying the responsible actor (public authority) and waiting in queues. At the same time, the percentage of users who acquaint with the Internet has been remarkably increased (Internet World Stats, 2005). These two facts motivate many governmental organizations to proceed with the provision of e-services via their Web sites. The ease and speed with which business transactions can be carried out over the Web has been a key driving force in the rapid growth and popularity of e-government, e-commerce, and e-business applications.

In this framework, the Web is emerging as the appropriate environment for business transactions and user-organization interactions. However, since it is a large collection of semi-structured and structured information sources, Web users often suffer from information overload. *Personalization* is considered as a popular solution in order to alleviate this problem and to customize the Web environment to users (Eirinaki & Vazirgiannis, 2003). Web personalization can be described, as any action that makes the Web experience of a user personalized to his or her needs and wishes. Principal elements of Web personalization include modeling of Web objects (pages) and subjects (users), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization.

In the remainder of this article, we present the way an e-government application can deploy Web mining techniques in order to support intelligent and personalized interactions with citizens. Specifically, we describe the tasks that typically comprise this process, illustrate the future trends, and discuss the open issues in the field.

BACKGROUND

The close relation between Web mining and Web personalization has become the stimulus for significant research work in the area (Borges & Levene, 1999; Cooley, 2000; Kosala & Blockeel, 2000; Madria, Bhowmick, Ng, & Lim, 1999). Web mining is a complete process and involves specific primary data mining tasks, namely data collection, data reprocessing, pattern discovery, and knowledge post-processing. Therefore, Web mining can be viewed as consisting of the following four tasks (Etzioni, 1996):

- **Information Retrieval—IR (Resource Discovery):** It deals with automatic retrieval of all relevant documents, while at the same time ensuring that the non relevant ones are fetched as few as possible. The IR process mainly deals with document representation, indexing, and searching. The process of retrieving the data that is either online or offline from the text sources available on the Web such as electronic newsletters, newsgroups, text contents of HTML documents obtained by removing HTML tags, and also the manual selection of Web resources. Here are also included text resources that originally were not accessible from the Web but are accessible now, such as online texts made for search purposes only, text databases, and so forth.
- **Information Extraction—IE (Selection and Pre-Processing):** Once the documents have been retrieved in the IR process, the challenge is to automatically extract knowledge and other required information without human interaction. IE is the task of identifying specific fragments of a single document that constitute its core semantic content and transforming them into useful information. These transformations could be either a kind of pre-processing such as removing stop words, stemming, etc. or a pre-processing aimed at obtaining the desired representation such as finding phrases in the training corpus, transforming the presentation to relational or first-order logic form, and so forth.
- **Generalization (Pattern Recognition and Machine Learning):** Discover general patterns at individual Web sites or across multiple sites. Machine learning or data mining techniques are used for the generalization. Most of the machine learning systems, deployed on the Web, learn more about the user's interest than the Web itself.
- **Analysis (Validation and Interpretation):** A data driven problem, which presumes that there is sufficient data available, so that potentially useful information can be extracted and analyzed. Humans also play an important role in the information or knowledge discovery process on the Web, since the Web is an interactive medium. This is especially important for validation and/or interpretation but under Etzioni's view (1996) of Web mining, "manual" (interactive, query triggered) knowledge discovery is excluded and thus the focus is placed on automatic data-triggered knowledge discovery.

Web mining refers to the overall process of discovering potentially useful and previously unknown information, knowledge, and patterns from Web data. In this sense, it implicitly covers the standard process of knowledge discovery in databases (KDD) and can be considered as a KDD extension applied to the Web (Markellos, Markellou, Rigou, & Sirmakessis, 2004a). Specifically, Web mining can be categorized into three areas of interest based on which part of the Web is mined:

- **Web Content Mining:** Focuses on the discovery/retrieval of useful information from Web contents/data/documents. Web content data consist of unstructured data (free texts), semi-structured data (HTML documents) and more structured data (data in tables, DB generated HTML pages)
- **Web Structure Mining:** Focuses on the structure of the hyperlinks within the Web as a whole (inter-document) with the purpose of discovering its underlying link structure. Web structure data consist of the Web site structure itself
- **Web Usage Mining:** Mines the secondary data derived from Web surfers' sessions or behaviors and focuses on techniques that could predict user behavior while the user interacts with the Web (Cooley, 2000). Web usage data can be server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the result of interactions

Recently, Web usage mining (Srivastava, Cooley, Deshpande, & Tan, 2000) has been proposed as an underlying approach for Web personalization (Mobasher, Cooley, & Srivastava, 2000). The goal of Web usage mining is to capture and model the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages or items that are frequently accessed by groups of users with common needs or interests. Such patterns can be used to better understand behavioral characteristics of visitors or user segments, improve the organization and structure of the site, and create a personalized experience for visitors by providing dynamic recommendations. In particular, techniques such as clustering, association rule mining, and navigational pattern mining that rely on online pattern discovery from user transactions can be used to improve the scalability of collaborative filtering when dealing with clickstream and e-government data.

WEB MINING TECHNIQUES IN E-PUBLIC SERVICES

For the implementation and successful operation of e-government, the proper design, which will be the basis in