Semantic Web Adaptation

Alexander Mikroyannidis
University of Manchester, UK

Babis Theodoulidis
University of Manchester, UK

INTRODUCTION

The rate of growth in the amount of information available in the World Wide Web has not been followed by similar advances in the way this information is organized and exploited. Web adaptation seeks to address this issue by transforming the topology of a Web site to help users in their browsing tasks. In this sense, Web usage mining techniques have been employed for years to study how the Web is used in order to make Web sites more user-friendly.

The Semantic Web is an ambitious initiative aiming to transform the Web to a well-organized source of information. In particular, apart from the unstructured information of today's Web, the Semantic Web will contain machine-processable metadata organized in ontologies. This will enhance the way we search the Web and can even allow for automatic reasoning on Web data with the use of software agents. Semantic Web adaptation brings traditional Web adaptation techniques into the new era of the Semantic Web. The idea is to enable the Semantic Web to be constantly aligned to the users' preferences. In order to achieve this, Web usage mining and text mining methodologies are employed for the semi-automatic construction and evolution of Web ontologies. This usage-driven evolution of Web ontologies, in parallel with Web topologies evolution, can bring the Semantic Web closer to the users' expectations.

BACKGROUND

Web Usage Mining

Web usage mining has a wide variety of applications. User profiles can be created for use in Web personalization. Information can also be extracted that details how a Web site can be reorganized to better facilitate users' navigation through it. In e-commerce Web sites, the results of Web usage mining can be used to improve sales. Analyzing user access patterns can also help when targeting advertisements to specific groups of users.

Srivastava, Cooley, Deshpande, and Tan (2000) divide Web usage mining into three stages:

i. Preprocessing
ii. Pattern discovery
iii. Pattern analysis

Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery. Usage preprocessing involves the identification of users and their visiting sessions. In order for this to be accomplished, several difficulties need to be overcome. For example, proxy servers hide the actual IP addresses of the machines that are using them, thus making user identification problematic. A user that uses more than one browser, even on the same machine, will appear as multiple users. Tracking repeat visitors can also be complex if a user uses different machines.

Content preprocessing consists of converting the text, image, scripts, or multimedia files into forms that are useful for the Web usage mining process. This often involves the application of content mining techniques, such as classification or clustering. For instance, a classification algorithm could be used to limit the discovered patterns to those that contain page views about a certain subject. Similar to the preprocessing of the site's content, structure preprocessing regards the extraction of the site's structure for use in the mining procedure. The hyperlinks of each Web page build the structure of the Web site. Most Web sites have nowadays an utterly dynamic topology, thus presenting a different structure to different users. This characteristic should be considered during the structure preprocessing phase.

The preprocessing stage is followed by the discovery of traversal patterns from the user access data. Traversal patterns reveal the way a user navigates through the site during each session. Clusters of users can be discovered through clustering of similar traversal patterns. Moreover, association rules can be applied to the pages accessed during a session, independent of their ordering. Examples of association rules that were extracted from an IBM analysis of the Web logs of the Official 1996 Olympics Web site (Elo-Dean & Viveros, 1997) are:

- 45% of the visitors who accessed a page about Indoor Volleyball also accessed a page on Handball.
- 59.7% of the visitors who accessed pages about Badminton and Diving also accessed a page about Table Tennis.

The percentages mentioned in both association rules are called confidence. Confidence can be defined as the number of transactions containing all of the items in a rule, divided by the number of transactions containing the rule antecedents (Cooley, Mobasher, & Srivastava, 1999). Additionally, temporal relationships among data items can be discovered, such as the following (Cooley, Mobasher, & Srivastava, 1997):

- 30% of clients who visited the '/company/products' page had done a search in Yahoo within the past week on keyword w.
- 60% of clients who placed an online order in the '/company/product1' page also placed an online order in the '/company/product4' page within 15 days.

Depending on the purpose of the mining, a traversal pattern may contain backward traversals. Backward traversals consist of references of pages earlier visited. Studying backward traversals can help discover missing hyperlinks, which if added will reduce these traversals, thus making navigation paths shorter and more convenient for the users. A pattern's accesses can also be restricted only to contiguous ones, which can be used for prefetching and caching purposes. A frequent pattern is maximal if does not contain any frequent subpatterns. This can reduce significantly the number of meaningful discovered patterns.

Pattern analysis is the last stage of the Web usage mining process. The patterns that have been produced are reviewed, and useful information is extracted from them. Knowledge query mechanisms, similar to structured query language (SQL), can be used to filter out the patterns. Another approach involves the use of data cubes and OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can also be utilized to highlight interesting trends in the data. Last but not least, content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

Numerous approaches to Web usage mining have been followed, targeted to a wide range of applications. Chen, Park, and Yu (1998) and Nanopoulos and Manolopoulos (2000) have introduced the concept of using the maximal forward references to break down user sessions into transactions for mining access patterns. Yang, Pan, and Chung (2001) have proposed an efficient hash-based method, HMFS, for discovering the maximal frequent itemsets. Spiliopoulou (1999) has presented an algorithm for building aggregating trees from Web logs, then mining the Web access patterns by MINT mining language. (Cooley et al. (1999) have provided a query language on top of external mining software for association rules and for sequential patterns. Another query language for extracting navigation patterns, called MiDAS, has been proposed by Buchner, Baumgarten, Anand, Mulvenna, and Hughes (1999). Xiao and Dunham (2001) have investigated techniques to discover frequently used contiguous sequences of page references, which they call maximal frequent sequences (MFS). They have also developed an algorithm called online adaptive traversal (OAT) pattern mining, to mine MFS. Xing and Shen (2004) have proposed two algorithms, user access matrix (UAM) and preferred navigation tree (PNT), for mining user preferred navigation patterns.

Web Adaptation

Several Web adaptation systems have been developed over the years, mainly based on Web usage mining techniques. The WebWatcher system (Joachims, Freitag, & Mitchell, 1997) suggests links that may interest a user, based on other users' online behaviour. The system is implemented in the form of a proxy server. Each user is asked, upon entering the site, what kind of