Mining Historical XML

Qiankun Zhao
Nanyang Technological University, Singapore

Sourav Saha Bhowmick
Nanyang Technological University, Singapore

INTRODUCTION

Nowadays the Web poses itself as the largest data repository ever available in the history of humankind (Reis et al., 2004). However, the availability of huge amount of Web data does not imply that users can get whatever they want more easily. On the contrary, the massive amount of data on the Web has overwhelmed their abilities to find the desired information. It has been claimed that 99% of the data reachable on the Web is useless to 99% of the users (Han & Kamber, 2000, pp. 436). That is, an individual may be interested in only a tiny fragment of the Web data. However, the huge and diverse properties of Web data do imply that Web data provides a rich and unprecedented data mining source.

Web mining was introduced to discover hidden knowledge from Web data and services automatically (Etzioni, 1996). According to the type of Web data, Web mining can be classified into three categories: *Web content mining*, *Web structure mining*, and *Web usage mining* (Madria et al., 1999). Web content mining is to extract patterns from online information such as HTML files, e-mails, or images (Dumais & Chen, 2000; Ester et al., 2002). Web structure mining is to analysis the link structures of Web data, which can be inter-links among different Web documents (Kleinberg 1998) or intra-links within individual Web document (Arasu & Hector, 2003; Lerman et al., 2004). Web usage mining is defined as to discover interesting usage patterns from the secondary data derived from the interaction of users while surfing the Web (Srivastava et al., 2000; Cooley, 2003).

Recently, XML is widely used as a standard for data exchanging in the Internet. Existing work on XML data mining includes frequent substructure mining (Inokuchi et al., 2000; Kuramochi & Karypis, 2001; Zaki, 2002, Yan & Han, 2003; Huan et al., 2003), classification (Zaki & Aggarwal, 2003; Huan et al., 2004), and association rule mining (Braga et al., 2002). As data in different domains can be represented as XML documents, XML data mining can be useful in many applications such as bioinformatics, chemistry, network analysis (Deshpande et al., 2003; Huan et al., 2004) and etcetera.

BACKGROUND

The historical XML mining research is largely inspired by two research communities: XML data mining and XML data change detection. The XML data mining community has looked at developing novel algorithms to mine snapshots of XML data. The database community has focused on detecting, representing, and querying changes to XML data.

Some of the initial work for XML data mining is based on the use of the XPath language as the main component to query XML documents (Braga et al., 2002; Braga et al., 2003). In Braga et al. (2002, 2003), the authors presented the XMINE operator, which is a tool developed to extract XML association rules for XML documents. The operator is based on XPath and inspired by the syntax of XQuery. It allows us to express complex mining tasks, compactly and intuitively. XMINE can be used to specify indifferently (and simultaneously) mining tasks both on the content and on the structure of the data, since the distinction in XML is slight.

Other works for XML data mining focus on extracting the frequent tree patterns from the structure of XML data such as TreeFinder (Termier et al., 2002) and TreeMiner (Zaki, 2002). TreeFinder uses an Inductive Logic Programming approach. Notice that TreeFinder cannot produce complete results. It may miss many frequent subtrees, especially when the support threshold is small or trees in the database have common node labels. TreeMiner can produce the complete results by using a novel vertical representation for fast subtree support counting.

Different from the above techniques, which focus on designing ad-hoc algorithms to extract structures that *occur frequently* in the snapshot data collections, historical XML mining focus on the sequence of changes among XML versions.

Considering the dynamic nature of XML data, many efforts have been directed into the research of change detection for XML data. XML TreeDiff (Curbera & Epstein, 1999) computes the difference between two XML documents using hash values and simple tree

comparison algorithm. XyDiff (Cobena et al., 2002) is proposed to detect changes of ordered XML documents. Besides *insertion*, *deletion*, and *updating*, XyDiff also support *move* operation. X-Diff (Wang et al., 2003) is designed to detect changes of unordered XML documents. In our historical XML mining, we extend the XML change detection techniques to discover hidden knowledge from the history of changes to XML data with data mining techniques.

MAIN THRUST

Overview

Consider the dynamic property of XML data and existing XML data mining research; it can be observed that the dynamic nature of XML leads to two challenging problems in XML data mining. First is the maintenance of the mining results for existing mining techniques. As the data source changes, new knowledge may be found and some old ones may not be valid. Second, is the discovery of novel knowledge hidden behind the historical changes, some of them are difficult or impossible to be discovered from snapshot data. In this paper, we focus on the second issue. That is, to discover novel hidden knowledge from the historical changes to XML data. Suppose there is a sequence of XML documents, which are different versions of the same documents. Then, following novel knowledge can be discovered. Note that, by no means we claim that the list is exhaustive. We use them as representatives for the various types of knowledge behind the history of changes.

- **Frequently changing/Frozen structures/contents:** Some parts of the structure or content change more frequently and significantly compared to other structures. Such structures and contents reflect the relatively more dynamic parts of the XML document. Frozen structures/contents represent the most stable part of the XML document. Identifying such structure is useful for various applications such as trend monitoring and change detection of very large XML documents.
- **Association rules:** Some structures/contents are associated in terms of their changes. The association rules imply the concurrence of changes among different parts of the XML document. Such knowledge can be used for XML change prediction, XML index maintenance, and XML based multimedia annotation.
- **Change patterns:** From the historical changes, one may observe that more and more nodes are inserted under certain substructures, while nodes are inserted and deleted frequently under others. Such change patterns can be critical for monitoring and predicting trends in e-commerce Web sites. They may imply certain underlying semantic meanings; and can be exploited by strategy makers.

Applications

Such novel knowledge can be useful in different applications, such as intelligent change detection for very large XML documents, Web usage mining, dynamic XML indexing, association rule mining, evolutionary pattern based classification, and etcetera. We only elaborate on the first two applications due to the limitation of space.

Suppose one can discover substructures that change frequently and those that do not (frozen structures), then he/she can use this knowledge to detect changes to relevant portions of the documents at different frequency based on their change patterns. That is, one can detect changes to frequently changing content and structure at a different frequency compared to structures that do not change frequently. Moreover, one may ignore frozen substructures during change detection, as most likely they are not going to change. As one of the major limitations of existing XML change detection systems (Cobena et al., 2002; Wang et al., 2003) is that they are not scalable for very large XML documents. Knowledge extracted from historical changes can be used to improve the scalability of XML change detection systems.

Recently, a lot of work has been done in Web usage mining. However, most of the existing works focus on snapshot Web usage data, while usage data is dynamic in real life. Knowledge hidden behind historical changes of Web usage data, which reflects how Web access patterns (WAP) change, is critical to adaptive Web, Web site maintenance, business intelligence, and etcetera. The Web usage data can be considered as a set of trees, which have the similar structures as XML documents. By partitioning Web usage data according to the user-defined calendar pattern, we can obtain a sequence of changes from the historical Web access patterns. From the changes, useful knowledge, such as how certain Web access patterns changed, which parts changes more frequently and which parts do not, can be extracted. Some preliminary results of mining the changes to historical Web access patterns have been shown in Zhao and Bhowmick (2004).

Research Issues

To the best of our knowledge, existing state-of-the-art XML (structure related) data mining techniques (Yan & Han, 2002; Yan & Han, 2003) cannot extract such novel knowledge. Even if we apply such techniques repeatedly