Discovering Knowledge from XML Documents

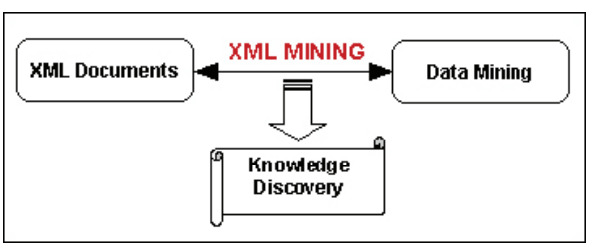
Richi Nayak
Queensland University of Technology, Australia

INTRODUCTION

XML is the new standard for information exchange and retrieval. An XML document has a schema that defines the data definition and structure of the XML document (Abiteboul et al., 2000). Due to the wide acceptance of XML, a number of techniques are required to retrieve and analyze the vast number of XML documents. Automatic deduction of the structure of XML documents for storing semi-structured data has been an active subject among researchers (Abiteboul et al., 2000; Green et al., 2002). A number of query languages for retrieving data from various XML data sources also has been developed (Abiteboul et al., 2000; W3c, 2004). The use of these query languages is limited (e.g., limited types of inputs and outputs, and users of these languages should know exactly what kinds of information are to be accessed). Data mining, on the other hand, allows the user to search out unknown facts, the information hidden behind the data. It also enables users to pose more complex queries (Dunham, 2003).

Figure 1 illustrates the idea of integrating data mining algorithms with XML documents to achieve knowledge discovery. For example, after identifying similarities among various XML documents, a mining technique can analyze links between tags occurring together within the documents. This may prove useful in the analysis of e-commerce Web documents recommending personalization of Web pages.

Figure 1. XML mining scheme



BACKGROUND: WHAT IS XML MINING?

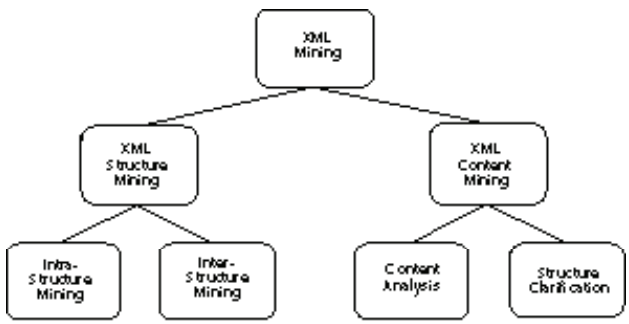
XML mining includes mining of structures as well as contents from XML documents, depicted in Figure 2 (Nayak et al., 2002). Element tags and their nesting therein dictate the structure of an XML document (Abiteboul et al., 2000). For example, the textual structure enclosed by <author>… </author> is used to describe the author tuple and its corresponding text in the document. Since XML provides a mechanism for tagging names with data, knowledge discovery on the semantics of the documents becomes easier for improving document retrieval on the Web. Mining of XML structure is essentially mining of schema including intrastructure mining, and interstructure mining.

Intrastructure Mining

Concerned with the structure within an XML document. Knowledge is discovered about the internal structure of XML documents in this type of mining. The following mining tasks can be applied.

The classification task of data mining maps a new XML document to a predefined class of documents. A schema is interpreted as a description of a class of XML documents. The classification procedure takes a collection of schemas as a training set and classifies new XML documents according to this training set.

Figure 2. A taxonomy of XML mining



The clustering task of data mining identifies similarities among various XML documents. A clustering algorithm takes a collection of schemas to group them together on the basis of self-similarity. These similarities are then used to generate new schema. As a generalization, the new schema is a superclass to the training set of schemas. This generated set of clustered schemas can now be used in classifying new schemas. The superclass schema also can be used in integration of heterogeneous XML documents for each application domain. This allows users to find, collect, filter, and manage information sources more effectively on the Internet.

The association data mining describes relationships between tags that tend to occur together in XML documents that can be useful in the future. By transforming the tree structure of XML into a pseudo-transaction, it becomes possible to generate rules of the form "if an XML document contains a <craft> tag, then 80% of the time it also will contain a <licence> tag." Such a rule then may be applied in determining the appropriate interpretation for homographic tags.

Interstructure Mining

Concerned with the structure between XML documents. Knowledge is discovered about the relationship between subjects, organizations, and nodes on the Web in this type of mining. The following mining tasks can be applied.

Clustering schemas involves identifying similar schemas. The clusters are used in defining hierarchies of schemas. The schema hierarchy overlaps instances on the Web, thus discovering authorities and hubs (Garofalakis et al. 1999). Creators of schema are identified as authorities, and creators of instances are hubs. Additional mining techniques are required to identify all instances of schema present on the Web. The following application of classification can identify the most likely places to mine for instances. Classification is applied with namespaces and URIs (Uniform Resource Identifiers). Having previously associated a set of schemas with a particular namespace or URI, this information is used to classify new XML documents originating from these places.

Content is the text between each start and end tag in XML documents. Mining for XML content is essentially mining for values (an instance of a relation), including content analysis and structural clarification.

Content Analysis

Concerned with analysing texts within XML documents. The following mining tasks can be applied to contents.

Classification is performed on XML content, labeling new XML content as belonging to a predefined class. To reduce the number of comparisons, pre-existing schemas classify the new document's schema. Then, only the instance classifications of the matching schemas need to be considered in classifying a new document.

Clustering on XML content identifies the potential for new classifications. Again, consideration of schemas leads to quicker clustering; similar schemas are likely to have a number of value sets. For example, all schemas concerning vehicles have a set of values representing cars, another set representing boats, and so forth. However, schemas that appear dissimilar may have similar content. Mining XML content inherits some problems faced in text mining and analysis. Synonymy and polysemy can cause difficulties, but the tags surrounding the content usually can help resolve ambiguities.

Structural Clarification

Concerned with distinguishing the similar structured documents based on contents. The following mining tasks can be performed.

Content provides support for alternate clustering of similar schemas. Two distinctly structured schemas may have document instances with identical content. Mining these avails new knowledge. Vice versa, schemas provide support for alternate clustering of content. Two XML documents with distinct content may be clustered together, given that their schemas are similar.

Content also may prove important in clustering schemas that appear different but have instances with similar content. Due to heterogeneity, the incidence of synonyms is increased. Are separate schemas actually describing the same thing, only with different terms? While thesauruses are vital, it is impossible for them to be exhaustive for the English language, let alone handle all languages. Conversely, schemas appearing similar actually are completely different, given homographs. The similarity of the content does not distinguish the semantic intention of the tags. Mining, in this case, pro-