

DATAMINING AND XML DOCUMENTS

Richi Nayak
School of Information System
Queensland University of Technology
GPO Campus, GPO Box 2434
Brisbane, QLD 4001, Australia

Rebecca Witt & Anton Tonev
School of Computer and Information Science
University of South Australia
Mawson Lakes
Adelaide, SA 5095, Australia

Abstract

XML is touted as the breakthrough in data exchange on the web. As XML material becomes more abundant, the ability to gain knowledge from XML sources decreases. Thus, there is a great need to apply data mining techniques to XML data. This paper suggests a taxonomy of XML mining as a stepping-stone to further XML mining research. This paper explicitly expresses the available classes of XML material. This paper also proposes surveys on a number of possible data mining techniques that can be applied on structure or content of XML documents.

Keywords: Data mining, XML, Taxonomy, XML structure mining, XML content mining

1 Introduction

The August 2001 estimate of web pages is 1.4 billion, increasing daily by 1 million.¹ Currently, the majority of these pages are in HyperText Markup Language (HTML). However, this is likely to change. As the corporate world becomes increasingly distributed, with a corresponding increase in data source heterogeneity, there is an obvious need for data exchange. HTML as a means of data exchange is inferior. Intended as a conveyance of technical reports, HTML tags are primarily for formatting markup.

Though some internal structural information is inferable from them (e.g. `<h1>` indicating an important information), HTML tags cannot give a clear indication of content. Moreover, the diversity of authorship guarantees no consistency between documents.

Here we have the motivation for eXtensible Markup Language (XML) - providing a markup language more conducive to data exchange. As semi-structured data, it is self-describing. An XML document has document type definitions (DTD) that define the structure of the document and what tags might be used to encode the document. Due to this advantage, XML will be the dominant format in a few years. If this prediction does prove true, then a number of techniques will be required to retrieve and analyse the vast amount of XML documents.

Data mining [6] will be essential for discovering new knowledge from many XML resources likely to arise in the next few years. Given the irony that humans produce far more data than they can ever analyse alone, the development of XML mining techniques must keep pace with development and implementation of XML itself.

¹<http://www.searchengineshowdown.com/>

This paper describes a classification of XML documents, and suggests taxonomy of XML mining. Discussing the application of mining techniques such as classification and clustering techniques, the paper then provides a summary of useful XML mining tools and techniques that can be applied on content or structure of XML documents.

1.1 A Classification of XML Documents

We use set notation to define the relationship between web pages and XML documents. Let all textual objects be the set T . Let web pages containing XML be called XML material for the remainder - be X , such that $X \subseteq T$. If D denotes the set of XML documents, then $D \subset X$, that is, XML material is not automatically classed as XML documents. Strictly, web pages are classed as XML documents only if they are well formed, $D = W$, where W is the set of well formed XML documents. Additionally, $V \subseteq W$, where V is the set of valid XML documents. Finally, this allows the definition of ill-formed XML material, given as $X \setminus W$.

Well-formed To be well-formed, a page's XML must have properly nested tags, unique attributes (per element), one or more elements, exactly one root element, plus a number of DTD-related constraints. Well-formed documents may have a DTD, but they do not conform to it.

Valid Valid XML documents are a subset of well-formed XML documents. To be valid, an XML document must additionally conform (at least) to an explicitly associated document type definition. A DTD for a document may include both internally and externally defined subsets (located within the same file or at different file, respectively). A DTD describes the grammar for an XML document

(allowing parsing). An XML document not conforming to its specified DTD is not valid, but may still be well formed.

Figure 1 illustrates the relationship between all set of XML material.

2 A Taxonomy of XML Mining

This section describes taxonomy of XML mining, specifically XML structure mining and XML content mining. This taxonomy is depicted in figure 2. For both categories, the application of data mining techniques, such as classification and clustering, are discussed. The type of XML material available for input to these procedures is also discussed.

2.1 XML Structure Mining

Element tags, and their nesting therein, dictate the structure of an XML document [3]. Mining for XML structure is essentially mining for schema. XML is semi-structured data, thus mining for XML structure is of interest. In this section, two further divisions in the taxonomy are presented: *intra-structure mining*, and *inter-structure mining*.

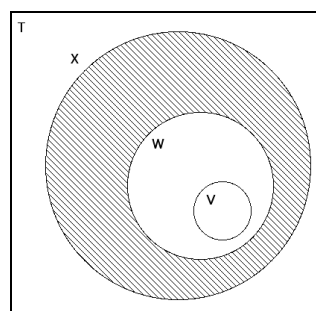


Figure 1 Venn diagram of XML
T: textual web objects,
X: XML material,
W: well-formed XML documents,
V: valid XML documents,
: ill-formed XML material

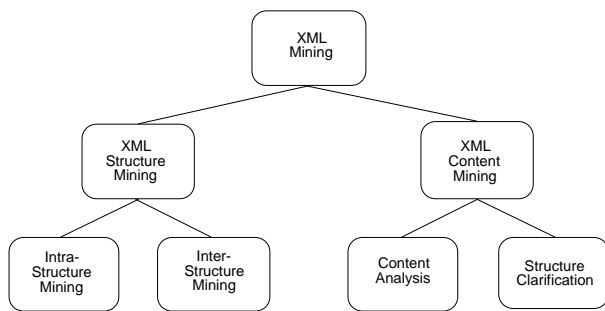


Figure 2 Taxonomy of XML Mining

Intra-structure Mining

Intra-structure mining is concerned with the structure within an XML document(s). Knowledge is discovered about the internal structure of XML documents, that is, their document type definitions.

Classification task of data mining can be applied to map a new XML document to a predefined class of documents. A DTD may be interpreted as a description of a class of XML documents. The classification procedure takes a collection of DTDs as a training set, and classifies new XML documents according to this training set of DTDs. This task is most easily performed on valid XML documents. With DTDs already defined for the new XML document, the classification can proceed by comparing the classification DTDs with the new DTDs.

For any XML document with an associated DTD, it should first be validated against it. It is important to distinguish between valid XML and well-formed XML within appropriate DTDs. For well-formed XML, an attempt is made to parse the documents according to the classification DTDs. A successfully parsed document is classified as an instance of the relevant DTD. Ill-formed XML may also be classified if enough of the document is parsed before an error occurs. Then the classification could be

used to 'rescue' any potential valuable information.

Clustering task of data mining can be applied for identifying similarities among various XML documents. A clustering algorithm takes a collection of DTDs to group them together on the basis of self-similarity. These similarities are then used to generate new DTD. As a generalisation, the new DTD is a superclass to the training set of DTDs.

Association rules discovery task of data mining can be applied to describe relationship between tags which tend to occur together in XML documents that can be useful in future. Transforming the tree structure of XML into 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 will also contain a <licence> tag." Such a rule may then be applied in determining the appropriate interpretation for a homographic tag.

Inter-structure Mining

Inter-structure mining is concerned with the structure between XML documents. Knowledge is discovered about the relationship between subjects, organizations and nodes on the Web.

Classification is applied with namespaces and URIs. Having previously associated a set of DTDs with a particular namespace or URI, this information is used to classify new XML documents originating from these places.

Clustering DTDs involves identifying similar DTDs. The clusters are used in defining hierarchies of DTDs. The DTD hierarchy overlaps instances on the web, thus discovering authorities and hubs [8]. Creators of DTDs are identified as authorities, and creators of instances are hubs. Additional mining techniques are required to identify all instances of DTD present on the web. The following application of

classification can identify the most likely place to mine for instances.

2.2 XML Content Mining

Content is the text between each start and end tag [3] in XML documents. Mining for XML content is essentially mining for values (an instance of a relation). The semi-structured nature of XML poses a challenge for content mining. However, a number of query languages designed for semi-structured data have been implemented. XML content mining can further be divided into two tasks: *content analysis* and *structural clarification*.

Content Analysis

Similar tasks are performed on XML documents as are performed on other text documents.

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

Clustering on XML content identifies the potential for new classifications. Again, consideration of DTDs will lead to quicker clustering: similar DTDs are likely to have a number of value sets. For example, all DTDs concerning vehicles will have a set of values representing cars, another set representing boats, etc. However, DTDs 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 can usually help resolve ambiguities.

Structure Clarification

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

Content may also prove important in clustering DTDs that appear different, but have instances with similar content. Due to heterogeneity, the incidence of synonymy is increased. Are separate DTDs 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 handling all languages. Or conversely, DTDs appearing similar are actually completely different, given homographs. For example:

```
<craft>boatbuilding</craft> and  
<craft>boat</craft>
```

Interpretation of the former is occupation, and of the latter vessel. The similarity of the content does not distinguish the semantic intention of the tags. Mining in this case can provide probabilities of a tag having a particular meaning, or a relationship between meaning and a URI.

3 Methods of Mining in XML Documents

After presenting taxonomy of XML mining, in this section we propose/survey a number of possible data mining techniques that can be applied on structure or content of XML documents.

3.1 XML Structure Mining

The main goal is to find a successful way of extracting structure from XML

documents, account for its characteristics and warehouse it, so data mining technique can be applied to look for interesting patterns. Based on the structure of the XML we can have two cases, extracting structure using DTD and extracting structure without the help of the DTD.

Extracting structure using DTD

Mining of structures from a well formed or valid document is straightforward because of the characteristics of the DTD. It is a schema and a grammar. This extracted DTD now can be easily used for creating a relational representation of the data. The structure can be presented as a table with attributes, which can accommodate the embedded data. If the hierarchy of the attributes is deeper, then database techniques such as adding more relations and foreign keys or/and normalization techniques could be used to accommodate the structure and the data.

Inferring structure without using DTD

Forming of structures from an ill formed document, the XML document is approached as Object Exchange Model (OEM) data by using the corresponding data graph to produce a most specific data guide. The data graph represents the interactions between the objects in a given data domain. When extracting a schema from a data graph the goal is to produce a most specific schema graph from the original graph. Two requirements have to be satisfied to be a most specific schema graph: (1) accuracy, meaning that every path in the new data graph occurs in the old data graph and every path in the old data graph appears in the new data graph; and (2) Concise, meaning that every path in the new data graph occurs exactly once.

A data graph satisfying those conditions is called a data guide or full

representative object [11,12]. This way of extracting schema is more general than using the DTD for a guide because (1) most of the XML documents do not have DTD, and/or (2) sometimes if they have, they do not confirm to it. Sometimes it is hard to construct a DTD confirming to the XML document because there is no consistent way in the number and type of attributes that are used to describe an object such as an employee of a company.

Inferring structure from existing query generated data

Some semistructured data are a result of queries. In such a case it is possible to derive the structure from the query that generated the data. In these cases, extracting a schema from the query is a better choice than extracting the schema from the data.

3.2 XML Content Mining

Before knowledge discovery in XML documents can occur, it is necessary to have some means of querying XML tags and content. A SQL based query can be performed to extract data from XML documents. There are a number of query languages, both specifically designed for XML, and those for semi-structured data in general.

Query Languages for Semi-structured Data

XML represents a subset of semi-structured data. Semi-structured data can be described by the grammar of ssd-expressions (semi-structured data expressions). The translation of XML to ssd-expression is easily automated [1]. Query languages for semistructured data exploit path expressions. In this way, data can be queried to an arbitrary depth. Path expressions themselves are elementary queries, with their results returned as a set of nodes. However, the ability to return results as semi-structured

data is required, which path expressions alone cannot do. Combining path expressions with SQL-style syntax provides greater flexibility in testing for equality, performing joins, and specifying the form of query results. Two such languages are *LoREL* [2] and *Unstructured Query Language (UnQL)* [7]. *UnQL* requires more precision, and is more reliant on path expressions.

Query Languages for XML

XML-QL, *XML-GL* and *XSL* are designed specifically for querying XML. *XML-QL* [8] brings together regular path expressions, SQL-style query techniques, and XML syntax. The great benefit is the construction of the result in XML, and thus transforming XML data from one DTD to another. *Extensible Stylesheet Language (XSL)* [13] is not implemented as a query language, but is intended as a tool for transforming XML to HTML. However, *XSL*'s 'select pattern' is a mechanism for information retrieval, and as such, is akin to a query [14]. *XML-GL* [4] is a graphical language for querying and restructuring XML documents.

4 Conclusion and Future Directions

Since there is an enormous amount of data on the web, and the exciting possibilities that data mining techniques are presenting, the idea behind the XML representation of data looks like a solution to a problem that those two earlier described phenomena. It is a challenging and exciting field with a lot of still open possibilities.

This paper explicitly expressed the available classes of XML material. It then taxonomised XML mining into two broad categories: XML structure mining, and content mining. Both of these categories were presented according to the data mining

tasks classification and clustering. This paper then proposes a number of possible data mining techniques that can be applied on the structure or content of XML documents.

Most of the goals addressed in the XML mining taxonomy need to be analysed for worthiness and practicality. For example, mining the web for instances of a particular DTD would be time consuming, if not impossible. However, web mining techniques listed in [8] should be easily catered to XML mining, as a subset of web mining after some modifications.

[9] identifies authorities and hubs by analysing the topology of hyperlinks. This can be extended to include external DTDs. The doctype declaration is an analogue to a hyperlink, an instance in effect linking back to its class definition. External DTDs thus become authorities. All instances of a DTD are hubs. [10] identifies subgraphs such as cliques and webs. A similar process may identify DTD hierarchies, as DTDs extend and link to each other. Enhanced hypertext categorization using hyperlinks [5] may be extended as enhanced XML categorisation using DTDs.

References

- [1] Abiteboul, S., Buneman, P. and Suciu, D. (2000) *Data on the Web: From Relations to Semistructured Data and XML*. California: Morgan Kaufmann.
- [2] Abiteboul, S., Quass, D., McHugh, J., Widom, J. and Weiner, J. (1997) 'The *LoREL* Query Language for Semistructured Data.' *Journal of Digital Libraries* 1(1):68-88.
- [3] Bray, T., Paoli, J. and Sperberg-Krcy, M. (1998) *Extensible Markup Language (XML) 1.0*: World Wide Web Consortium.
- [4] Ceri, S., Comai, S., Damiani, E., Fraternali, P., Paraboschi, S. and Tanca, L. (1999) 'XML-GL: A Graphical Language for

- Quering and Restructuring XML Documents.' *8th International WWW Conference*, Toronto.
- [5] Chakrabarti, S, Dom, Band Indyk, P (1998) 'Enhanced Hypertext Categorization Using Hyperlinks.' *Proceedings of the ACM SIGMOD Conference on Management of Data*.
- [6] Fayyad, UM, Piatetsky-Shapiro, G and Smyth, P (1995) 'From Data Mining to Knowledge Discovery: An Overview.' *Advances in Knowledge Discovery and Data Mining*. Fayyad, UM, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R. Menlo Park, AAAI Press : 1-34.
- [7] Fernandez, M and Suciu, D 'UNQL: A Query Language for Web Sites.' <http://www.cs.huij.ac.il/~yarivi/unql-htom.html>.
- [8] Garofalakis, M, Rastogi, R, Seshadri, S and Shim, K (1999) 'Data Mining and the Web: Past, Present and Future.' *Proceedings of the second international workshop on web information and data management*, Kansas City, USA.
- [9] Kleinberg, J (1997) 'Authoritative Sources in a Hyperlinked Environment.' *ACM-SIAM Symposium on discrete algorithms*.
- [10] Kumar, R, Raghavan, P, Rajagopalan, S and Tomkins, A (1999) 'Extracting Large-Scale Knowledge Bases from the Web.' *Proceedings of the International Conference on Very Large Data Bases*, Edinburgh, Scotland.
- [11] Nestorov S., Ullman J., Weiner J., Chawathe S. *Representative objects: Concise Representation of Semistructured, Hierarchical data*, IEEE 1999
- [12] Wang Q., Yu Xu Jeffrey, Wong K. *Approximate Graph Scheme Extraction for Semi-structured data*, Advances in database technology proceedings 2000.
- [13] W3C (1998) 'The Query Language Position Paper of the XSL Working Group.' *Proceedings of the Query Language Workshop*, Massachusetts.
- [14] W3C (2000) *XSL Specification*. Working Draft 18 Oct. <http://www.w3c.org/TR/WD-xsl>.