Data Mining and XML: Current and Future Issues[*]

A. G. Büchner[1], M. Baumgarten[2], M. D. Mulvenna[1], R. Böhm[2], S. S. Anand[1]

[1] MINEit Software Ltd, 5A Edgewater Business Park, Belfast, BT3 9JQ, UK

[2] Northern Ireland Knowledge Engineering Laboratory, University of Ulster, UK

alex@mineit.com

Abstract

The paper describes potential synergies between data mining and XML, which include the representation of discovered data mining knowledge, knowledge discovery from XML documents, XML-based data preparation, and XML-based domain knowledge. Each category is viewed from a theoretical as well as a practical point of view.

1. Introduction

Data Mining has been one of buzzwords in the business intelligence research and industry communities over the last few years. Similarly, XML [8] has drawn much attention across industries as yet another panacea for representing and exchanging information on the Internet. The objective of this paper is to investigate the synergy of the two fields and to discuss current as well as future issues[1].

The area has been sub-divided into four main areas, which are proposed separately. These are

- Representing discovered data mining knowledge;

- Knowledge Discovery from XML documents;

- XML-based data preparation; and

- XML-based domain knowledge.

For each topic, some background is given and potential application scenarios are provided. Since XML is mainly a web-related standard, the chosen scenarios have been kept as close to the web mining landscape as possible.

2. Representing Discovered Data Mining Knowledge

2.1 The General Idea

In recent years a change in knowledge discovery has occurred, which can be divided into three generations. First was mainly concerned with the development of more powerful data mining algorithms that discover better patterns, achieve higher accuracy, and scale better on large data sets (performance). The second generation tackled the knowledge discovery life cycle, which includes human resource identification, problem specification, data prospecting, domain knowledge incorporation, methodology identification, data pre-processing, pattern discovery, and knowledge post-processing [4]. The third and current generation is tackling the interchange of discovered knowledge among compliant applications, which requires the specification of a commonly accepted representation. This endeavor has been approached by PMML, which is now described.

2.2 Predictive Model Markup Language

The Data Mining Group (DMG) has defined the Predictive Model Markup Language (PMML) as "… an XML-based language which provides a quick and easy way for companies to define predictive models and share models between compliant vendors' applications". OLE/DM by Microsoft Corp. supports the standard in that it uses PMML as its XML output format.

"A PMML document provides a non-procedural definition of fully trained or parameterized analytic models with sufficient information for an application to deploy them. By parsing the PMML using any standard XML parser the application can determine the types of data input to and output from the models, the detailed forms of the models, and how, in terms of standard data mining terminology, to interpret their results." [6]
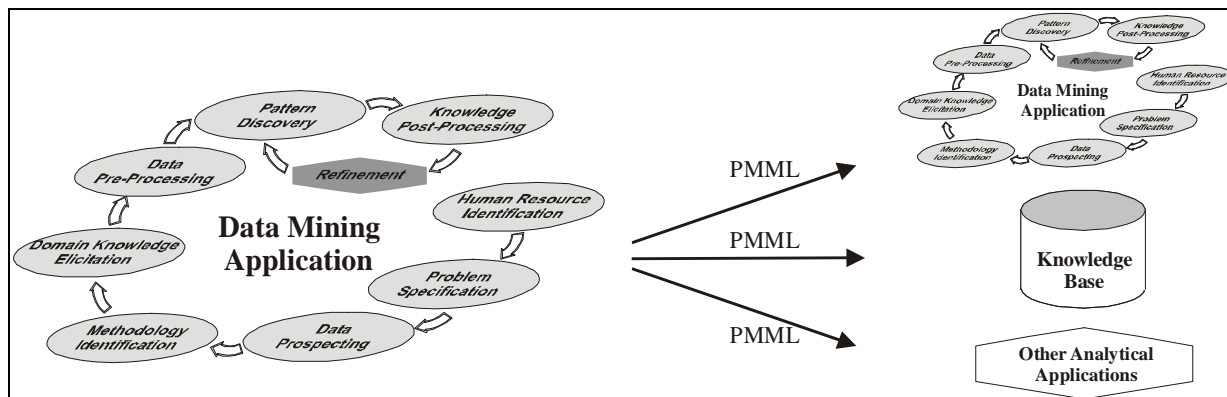
[1] It is assumed that the reader is familiar with the fundamentals of data mining and of XML.

Figure 1. Deploying PMML Output

Version 1.0 of the PMML standard provides a small set of DTDs that specify the entities and attributes for documenting decision tree and multinomial logistic regression models. Currently, further specifications are designed which include DTDs for clustering, neural networks, association rules, and others. They will be made available in Version 1.1.

2.3 Applications

Potential applications of PMML are manifoldand are shownpicted in Figure 1.

The most obvious application is the interchange of discovered patterns among conforming data mining systems. For example, it will be possible to use a web mining product to discover page segments and a separate sequence detection tool to find browsing behavior or click streams. Both result sets can then be exported to PMML and a third application that is capable of importing both models can utilise the results.

The exporting application does not necessarily have to be another knowledge discovery system. Any other analytical application, such as a spreadsheet, or reporting tool has the potential to employ exported PMML.

Since PMML allows the backtracking of model creation, which includes the model source, the used data, as well as specified algorithmic parameters, distributed data mining architectures are much more straightforward to build. The Kensington Project [9] is one candidate of such a system.

PMML output can further be used as filters for further data mining exercises. In Section 5, XML-based domain knowledge is presented in more detail.

Another promising application type exists in the area of recommendation systems. Storing PMML code in a knowledge base allows the usage from within a web server recommendation engine in order to provide personalized content.

3. Knowledge Discovery from XML Documents

3.1 Some Background

Data Mining is usually carried out on structured data, which is hosted in text files or, ideally, in database systems or data warehouse. However, a more recent discipline that has approached is the mining of unstructured and semi-structured data (see [11] and [10] for overviews, respectively), such as HTML documents (also known as web content mining). The challenge of such exercises supersede the ones of mining structured data in that it adds problems in the field of semantics.

Since XMLs main purpose is to add semantic aspects to web content, knowledge discovery in such documents may be easier to carry out. The added meta information simplifies the pre-processing of text that produces a document-based on concept-based intermediate form, which can then be used for further data mining activities.

3.2 Applications

Most applications in the area of knowledge discovery from XML documents will be strongly Internet-related.

Search engines have much greater opportunities to use context-sensitive information. For example, when searching for a particular key word in a special field, it is possible to reduce the result set by considering the domain-related semantics, but also to broaden the search by including other terms, which are closely related to the searched item. Machine learning mechanisms that support

the search can be extended in order to consider the given semantics more explicitly.

Personalized content provision has been a popular feature on information provider sites, in particular with the emerging WAP technology becoming more popular. This can range from news to electronic television guides. The discovery of relevant content can either be based on previous behavior (ideally stored in PMML) or on a profile conforming to P3P for example, which has been provided by the user (expressed in XML).

Price watchers that provide users with comparisons of prices across electronic shopping malls have mainly been criticized for their inaccuracy. With XML-tagged data, it is more likely to discover products and prices that match a user's needs. This example can be extended to other consumer-2-business electronic commerce scenarios. A related business-2-business application can be thought of as when different suppliers have agreed on a set of DTDs, going far beyond the Dublin Core standards.

4. XML-based Data Preparation

4.1 General Data Preparation Steps

Data preparation is one of the most time-consuming, expensive and tedious tasks in a data mining project [1] (sometimes as much as 80 per cent of the overall effort). This becomes even more challenging when using data from internet servers, since the information is usually stored in quasi-machine readable form [2, 5]. Thus, any simplification of this step can reduce the effort of a knowledge discovery exercise enormously. In order to demonstrate the potential of XML-based data preparation, two selected data cleansing scenarios with and without XML are described. The first setting deals with the problem of data enrichment, the second with the pre-processing of web log files.

Data enrichment involves the linkage of (usually) external information to the data that is being mined. The two main problems are the identification of matching field names and the matching of field values with different semantics, for example, different currencies.

The pre-processing of web log files aims at creating a canonical form, which contains the information that can be stored by a web server. Fields that are kept track of are hostname, document, referrer, date & time, etc. Since there are around a dozen logfile formats, plus the possibility by some web servers of creating user-defined log files, the data pre-processing can be a challenging task to perform.

4.2 Data Preparation without XML

Assuming a comma separated text file with no field information containing customer records and a database, which contains third party demographic and sociographic data, it can be challenging to find matching fields. Zip codes, after being identified, have to be matched, accordingly. From the database schema it is not possible to know in what unit some information is stored. For instance, currencies are not represented explicitly, or economic measures such as unemployment can have different definitions. Ignoring such information might lead to skewed knowledge discovery, which requires adding semantics manually.

When reading in web log files, which usually use different delimiters for different fields, the interpretation can be rather tedious. Some log files only provide the most essential fields, such as hostname, time and document, others store a range of up to 20 fields, including referrers, cookies, and so forth. Some fields cannot uniquely be identified, for instance bytes sent and bytes read. Date, time and time zone information is not represented equally on different log files, which becomes even more difficult when servers are scattered across countries. Finally, some fields, such as query strings, contain sub-information, like the keywords typed into a search engine.

4.3 Data Preparation with XML

Assuming the used data is stored in XML or at least the field information is represented in the markup language and the database schema has been exported to XML (for instance through OLE/DB), the integration of the two disparate data sources becomes much more transparent. Field names can be matched more easily and semantic conflicts may be described explicitly.

Having an agreed DTD for log files, which would contain the about 20 mentioned fields as well as some additional information, such as server name, server type, operating system, etc., the transposition from the machine-readable file to a canonical form which can be used as data mining input would be a straightforward exercise.

5. XML-based Domain Knowledge

5.1 The Concept of Domain Knowledge

So far, this article has dealt with representing data and knowledge using some form of XML; another ingredient for most data mining exercises is that of domain knowledge.

Domain knowledge can be utilised in a number of ways. It can be used for making patterns more visible, for constraining the search space, for discovering more accurate knowledge, and for filtering out uninteresting knowledge [1]. There are numerous types of domain knowledge, where the most popular ones are taxonomies, which encompass bandings (ranges), concept hierarchies and network models, constraints (also known as attribute-relationship rules), previously discovered knowledge, and user preferences [3].

No matter what type of domain knowledge has to be represented, it can always being brought down to a specific format, which allows the specification of a document template. This can range from simple parameters, such as a from-to range, to complex user preferences in the form of a profile.

5.2 Applications

The employment of XML-based domain knowledge is threefold. First is within or among data mining systems, second in the context of user profiling, and third is using already defined XML applications.

Taxonomies and constraints, not matter what form, can easily be represented in XML; a standardized approach does not exist yet. Previously discovered knowledge can be formulated in PMML, as presented in Section 2. The specific algorithm has to be capable of interpreting PMML patterns as domain knowledge or provide a mechanism to map PMML onto its existing domain knowledge format.

User profiling biases the discovery of unknown patterns in data with the objective to filter out unwanted knowledge. There have been numerous suggestions for profiling schemes, each of which could straightforwardly be represented in XML. The best known example is the Open Profiling Standard [7], which allows the user to specify preferences s/he is willing to provide online, and gives a personalization mechanism priorities which it can take into account.

There are numerous XML applications, that is supersets of XML, which describe a specific domain, similar to PMML. Examples are the Chemical Markup Language, the Mathematical Markup Language, and the Vector Markup Language. If a data mining exercise is to be carried out in such a domain the specific application can be used to specify the domain knowledge that has to be incorporated. For instance, the Channel Definition Format, which defines channels that are used to upload information to a user. Specifying domain knowledge in the context of a personalization exercise allows the data mining mechanism that is used for the discovery of personalized content, to consider such input.

6. Conclusions

It has been shown that the synergy of XML and data mining has great potential in manifold areas, both for the research community as well as for their industrial counterparts[2]. However, there are quite a few issues that might cause concerns in the near future.

Firstly, XML documents are (much) bigger than their textual equivalents. This is true for both, data represented in XML as well as discovered knowledge. The trade-off between additional semantic information on the one side and performance on the other side needs to be addressed.

Secondly, there has to be agreement among the participants in a specific area. While this issue seems to have been successfully addressed in the PMML consortium (most data mining vendors, Microsoft and IBM, as well as academic institutions have joined up), other areas have failed to do so. Either no attempts have been made in order to define an agreeable ontology or multiple approaches exist (the mentioned XML applications and some others being exceptions).

Thirdly, XML is still maturing which results in follow-up versions and related standards on a regular basis. Also, the technology for dealing with XML and related documents is still in its infancy (for instance, neither of the latest popular web browsers is able to check an XML document against a given DTD properly). Furthermore, many applications do not yet support XML; to date no existing commercial data mining products fully supports PMML.

However, considering all the potentials of the XML and data mining synergy, as well as the possibility to tackle the outlined drawbacks, a great future can be foreseen for the two mutual beneficial technologies.

7. References

[1] S.S. Anand, A.G. Büchner. Decision Support Using Data Mining, Financial Times Pitman Publishers, 1998.

[2] A.G. Büchner, M.D. Mulvenna. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining, *ACM SIGMOD Record*, 27(4):54-61, 1998.

---

2 Various XML-related subjects, although relevant to the presented subject, have not considered in here on purpose, for example, XML Query, XML Schema, style sheets, and XSL.

[3] A.G. Büchner, J.G. Hughes, D.A. Bell. Contextual Data and Domain Knowledge for Incorporation in Knowledge Discovery Systems, *Proc. 2nd Int'l and Interdisciplinary Conf. on Modeling and Using Context (CONTEXT-99)*, Trento, Italy, Lecture Notes in Artificial Intelligence, Vol. 1688, Springer-Verlag, pp. 447-450, 1999.

[4] A.G. Büchner, M.D. Mulvenna, S.S. Anand, J.G. Hughes. An Internet-enabled Knowledge Discovery Process (award paper), *Proc. 9th Int'l. Database Conf.*, Hong Kong, 13-27, 1999.

[5] R. Cooley, B. Mobasher, J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, 1(1), 1999.

[6] R. Cover. The XML Cover Pages: Predictive Model Markup Language (PMML), http://www.oasis-open.org/cover/pmml.html, 1999.

[7] http://www.w3.org/TR/NOTE-OPS-FrameWork.html

[8] http://www.w3.org/XML/

[9] Kensington. The Open Infrastructure for Enterprise Data Mining, Imperial College of Science, Technology and Medicine, http://kensington.doc.ic.ac.uk/, 1999.

[10] L. Singh, B. Chen, R. Haight, P. Scheuermann, K. Aoki. A Robust System Architecture for Mining Semi-Structured Data, *Proc. 4th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 329-333, New York City, 1998.

[11] A.-H. Tan. Text Mining: The state of the art and the challenges. *Proc. PAKDD'99 workshop on Knowledge Disocovery from Advanced Databases*, Beijing, p65-70, 1999.