International Journal of Information Technology and Knowledge Management
January-June 2012, Volume 5, No. 1, pp. 129-134

XML BASED INFORMATION MANAGEMENT IN HEALTHCARE, DATA MINING AND DISTRIBUTED SYSTEMS

Sudesh Kumar[1], and Deepak Sinwar[2]

ABSTRACT: XML Based Information Management is an emerging standard which has gain its popularity because of its various features like transfer of data from one system to another without having a common environment between them and also due to elimination of its dependency on relational databases like Oracle, DB2, Sybase etc. Now days, various industries have developed their own formats for exchanging data to support the environment among professionals to share information in distributed systems in a meaningful way. This paper will present some of the components of XML as well as some potential applications including healthcare, data mining, and distributed systems.

Keywords: Healthcare, Warehousing, Exchange, Dataset, Resource Scheduling

1. INTRODUCTION

XML provides a mechanism for the representation of hierarchical data in text format [18]. By representing data as a simple stream of text, some space efficiency is sacrificed in exchange for ease of transmission and retrieval. Another benefit of representing data in this way is that existing Internet protocols can be used to exchange the data between systems. For example, standard HTTP (Hypertext Transfer Protocol) servers can be used to serve up data requests between divisions. The data can be represented as static documents available on the HTTP server or (more likely) generated dynamically from a CGI (Computer Gateway Interface) script on the server. Data can be made available through the standard security mechanisms of the Internet, namely Secure Sockets Layer (SSL) via HTTPS.

Now days, the union of different technologies is the main challenge into D2H2 (Distributed Diagnosis and Home Healthcare) [14]. The market offers many good solutions to implement an efficient healthcare environment. Systems used at the present days are efficient although they have a fault: they only work with their own technology. When a user buys a new technology to improve his standard of living or to monitor a person, he needs to acquire all the devices which must be compatible with the technology chosen. But plenty of consortiums have been emerged along with XML in recent years to lesser down the possibilities of using the existing technology. For example, The 'MedBiquitous' consortium having a standard that allows licensing boards, certifying boards, education certifiers, and research databases to use a common language for data exchange within their own domains and with one another The exchange of data about healthcare professionals is essential to protecting public safety and enabling the practice and education of healthcare professionals. Yet often the organizations that collect and maintain data about healthcare professionals have difficulty in integrating data from multiple sources effectively and efficiently. Standards are essential to facilitate data integration and credentialing. The Healthcare Professional Profile technology standard allows data collectors to keep their data up-to-date more quickly and effectively. The XML format can also be used for credentialing Web services, potentially aggregating data from several umbrella organizations.

On the other hand Data Mining is one of the major application areas of XML, where data from heterogenous sources is pre-processed using ETL (Extract Transform and Load) techniques. If we look data mining as a revolutionary concept towards knowledge mining then a framework called KDD (Knowledge Discovery from Databases) is used. The KDD itself is a combination of cleaning, selection and machine learning algorithms introduced by data mining, because data mining is a step towards KDD. We will see the applications of XML in data mining in terms of ARM (Association Rule Mining), which itself produces rules for optimizing the performance of any system whether it is an information system or it is a kind of market-basket analysis from transaction dataset. We will use the term dataset, database and data-set interchangeable throughout this paper.

Distributed systems and information mediation addresses the problem of efficiently querying large numbers of text documents, using parallel processing methods [6]. The optimization criteria are somewhat different from those used in querying heterogeneous databases, largely because

[1] Associate Professor, Department of Information Technology, BRCM CET, Bahal (Bhiwani, Haryana, India,
E-mail: sudeshjakhar@gmail.com,
[2] Assistant Professor Department of Information Technology, BRCM CET, Bahal (Bhiwani, Haryana, INDIA)
E-mail: deepak.sinwar@gmail.com

the extraction of ontological information from documents is the dominant component of query execution time. On the other hand some researchers have been proposed some novel approaches for pervasive applications that communicate via an XML based distributed virtual shared information space [11].

The rest of the paper is organized as follows. In the next section we will review some work related with our work. Section III, IV and V will elaborate the details about the application areas introduced in first section, and section VI concludes the paper with possible suggestions regarding future work.

2. RELATED WORK

As an emerging technology, XML is gaining momentum as a potential solution to many of healthcare's most challenging problems. It is an exciting development in an industry that is under constant pressure to do more with fewer resources [18]. With this aim to effectively processing of distributed healthcare records Shammary and Khalil'10 comes with a new XML-aware compression technique for improving performance of the aforesaid area over the hospital networks. They proposed depth-first and breadth-first traversals to generate a textual expression of the medical XML message tree as they remove the redundant XML tags from the entire XML message. At the same time, fixed-length and Huffman as a variable-length encodings are developed in order to encode the status medical XML message tree. Now days, SOAP (Simple Object Access Protocol) is used in many hospital networks as a communication protocol for web services [19], but having some limitations. In healthcare SOAP, there are mainly five components including General Practitioner, Hospital Server, Patient Server, Patient Database and the Internet. However, SOAP Web services inherit the drawbacks of XML as Web messages which are larger than the real payload of services causing high network traffic over the net. Therefore, medical Web services often suffer congestion and bottlenecks as a result of the large size of the medical Web requests and responses sent and received over the Internet [19]. Whereas in case of data mining using XML various researchers [7, 10, 16, 17, 25, 26] have proved that XML can easily be used to mine knowledge from large databases.

3. XML IN DISTRIBUTED HEALTHCARE SYSTEMS

Communication is one of the crucial challenges in distributive systems, but due to the availability of a de-facto standard (i.e. XML) now days it is possible to efficiently communicate within the distributed environment. Moreno et al.'07 proposes an approach for communication between different technologies. They use XML to define the communication frames, stored data and the accessibility. They also proposes the ways to access to the information from anywhere using criteria of security, accessibility and the possibility of defining models of actuation before critical situations. The communication process of distributed systems begins in the sensors; the information collected from the patient is sent to the central core. The system analyses and stores the data so that the information can be available for everybody; implementing an interface with different methods of access. The information used in the environment is private; therefore the model has to implement cryptographic techniques to avoid any non authorized access or modification. Any frame, which is carrying the patient's information, is encrypted before being sent to the central core, in order to avoid being sniffed. When the central core receives the frame it is decrypted and then it is translated into a XML frame. With the new frame, the system analyses the information and verifies if it comes from the sensor or not [14]. The system includes some control process that makes possible the definition of actuation models, oriented to danger or risk situations. Many new possibilities appears from this environment; the patient can be monitored every time, including when is out of home. It is possible to use portable devices (e.g. a PDA or mobile) in the exterior, taking information and synchronizing with the central core when the patient arrives to home. Furthermore the system can warn someone automatically in an emergency. Whereas Maldonado et al.'03 describes the architecture, design and implementation of the PANGEA system, which allows healthcare professionals to access patient information stored in heterogeneous autonomous information systems through a set of formal aggregates of health data based on the European pre-standard of Healthcare Record Architecture ENV13606 from CEN/TC251. ENV13606 is also used as canonical model for the representation of healthcare information, therefore the overall system can also be considered as a system for publishing legacy relational data as XML-Electronic Healthcare Records compliant with ENV13606. The main components of the PANEGA system are:

(A) The metadata Server

It is the module that is in charge of managing the system's data dictionary. It manages an object-oriented database that basically contains the archetypes definitions, the underlying databases schemas, the archetypes-databases schemas mappings, information about the location of patient's social demographic data and general information about the underlying databases. Two visual tools have been developed to assist in the creation and management of metadata:

The archetype editor: It facilitates the edition of archetypes, which allows the creation of new ones from scratch or the reuse of the existing ones, it validates their

correctness, it allows their classification into groups in order to make the search easier, define the mappings with the component databases schemas and it controls the versioning (all prior versions are kept for legal reasons). The schemata manager: It allows the retrieval and caching of the underlying database schemas, the enrichment of the schemas by defining new relationships between attributes (foreign keys), define where the social demographic data about the patients, such as name, surname, address, SSN, date of birth etc., are located in order to allow the matching of patient identifiers and define which data from the underlying database is shared.

(B) The Electronic Healthcare Record Server (EHR Server)

It is the core of the whole system. It is layered between the client applications and the data repositories. This server retrieves, by request, all the relevant patient information wherever it is located and presents back the information as a XML document compliant with ENV13606. Client applications ask for health information about a particular patient as one or more instances of any archetype defined in the data dictionary.

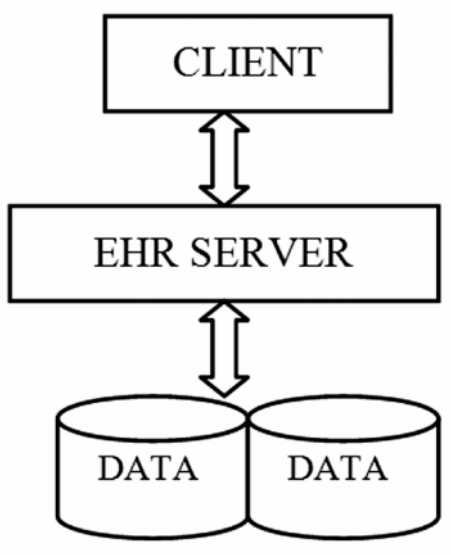

Figure 1: Electronic Healthcare Record Server (EHR Server)

The EHR server obtains the definition and mapping specifications of the requested archetype from the metadata server. Afterwards, it builds and populates, by interpreting the mapping specifications, the XML documents that contain the healthcare record extract. The EHR server offers a set of web services that can be used by client applications. Basically, a web service is an interface that describes a collection of operations that are network accessible through standardized XML messaging. The protocol designed to manage the XML messaging is called SOAP (Simple Object Access Protocol). This protocol defines a standard structure, encoding rules and associations to transport XML documents through other protocols such as HTTP, SMTP, etc. SOAP allows a high level of interoperation between heterogeneous applications as described in Section II. Therefore, it suits perfectly to accomplish the desirable requirement that between-organization communications needs to be achieved, ideally using the same technological solution as for intra-organizational communication.

4. SEARCHING PATTERNS USING XML

Pattern searching is the first step towards Association Rule Mining, which are searched on the basis of information domain available in transaction dataset. Association rule mining is one of the data mining techniques in which we find out interestingness and co-relationships among large set of data items. Association rules are generally formed by extracting large dataset for frequent patterns and then generating rules from these frequent item sets. A typical and widely-used example of association rule mining is Market Basket Analysis [1]. According to Abazeed et al.'09 the web data (XML data) is different from relational data in term of structure; relational data is flat and have a regular structure and is govern by data types while XML structure vary and consist of tags and some user defined tags . In order to mine XML data a data preparation step should take place, data preparation depends on the nature of the XML document and the type of transformation need to be done on the XML document in order to access it and mine it. Mining XML can be categorized into two ways:

1. Indirect mining which means preprocessing of the XML documents and transform them into a different structure (flat file for example) ,then apply the mining algorithm on the data after transformation in a relational format. And produce the result in XML format which is called post processing.
2. Direct mining which means mine the XML file without preprocessing or post processing, the XML file will be used as an input for the mining algorithm and the results can be displayed in XML format as well.

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, such as HTML documents (also known as web content mining). The challenge of such exercises supersedes 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. Various researchers [20, 22,25], have elaborated different applications of XML in data mining. Win et al.'05 have proposed an easy way to represent transaction data in the form of XML.

Table 1
Sample Transaction Dataset

TID (Transaction ID)	Items
1	A,B
2	A,D
3	A
4	B,C

We will use the data available in table 1 as a synthetic transaction data set to represent it in the form of XML as follows:

```
<transactions>
    <transaction id="1">
            <items>
            <item> A </item>
            <item> B </item>
    </items>
</transaction>
<transaction id="2">
    <items>
            <item> A </item>
            <item> D </item>
    </items>
</transaction>
<transaction id="3">
    <items>
            <item> A </item>
    </items>
</transaction>
<transaction id="4">
            <items>
            <item> B </item>
            <item> C </item>
            </items>
    </transaction>
</transactions>
```

Figure 2: Transaction Document (Transactions.xml)

We can then use the above XML document to process XQuery functions to search frequent patterns towards association rule mining.

Li et al.'07 have extended the notion of associated items to XML terminal-elements to present associations among tree-structured items; terminal-element is the element without sub element. According to them, a transaction in XML context is XML fragment that define the context in which the items must be counted. In other words, the transaction is a sub tree, and the items are the leaf nodes in the sub tree; they use the root node of sub tree to identify a transaction and use leaf node in the sub tree to identify an item. We define an XML association rule as an implication of the form $X \Rightarrow Y$, $X \subseteq I$, $Y \subseteq I$, and $X \neq Y$, where I is a set of terminal-elements (tree-structured items) [10].

5. INFORMATION MANAGEMENT

XML documents do not ordinarily contain information about how the data contained within the data should be rendered. The graphical rendering of an XML document can be achieved by the use of an "eXtensible Style Language" (XSL) style sheet. More generally, XSL style sheets can be used to not only render XML information graphically but can also be used to transform XML documents into other formats. Any number of XSL style sheets can be associated with an XML document. For example given that some management information is represented in the form of an XML document, it would be possible to write a web-based application that graphically represented the management object in a variety of ways simply by applying a number of different style sheets to the document. Cleveland' 08 has introduced IEM (Information Exchange Modelling) which may contain at least one of the following items:

1. The "Use Case" (as defined by UML) or the logical flow of information for each type of exchange of data between each pair of applications or systems.
2. Data object "Class" definitions, which define the name of each class of data, the class attributes, and the operations that can be performed on the class.
3. Interaction Event Sequence Diagrams, as necessary for the more complex interactions, to clarify messaging processes, such as what events may trigger an exchange, what types of messages are to be used, which classes of data are included, etc.
4. Interface Diagrams, describing the blocks of data classes to be transferred upon each type of event, as well as methods used to compress the data transfers (e.g. use of well-defined data sets with report by exception, zip files, named data items with report by exception, etc.)

5. Interface specifications, to define the actual interface details, including the type of interface, such as standardized Corba API, EJB, SQL query, ftp (flat file transfer), Oracle forms, or CSV files, etc.

6. Data specifications, to define the actual data and formats. This last step moves from the modelling of these data exchanges to the actual implementation-specific bits and bytes.

The new Internet Web technology - extensible Markup Language (XML) is one powerful tool to actually define and implement the Information Exchange Models [5]. Whereas, on the other hand, XML can also be used for modeling information structure of Electronic Information System [15].

6. CONCLUSION

XML is generally a powerful standard which can be used in different areas like information modelling, warehousing, transfer of data from one system to another and many more. With this aim we present the power of XML by considering its contribution in three major areas viz. Distributed Healthcare Management, Pattern searching and Information Management. The work may be extended to analyse the performance of XML by modelling applications through XML large and real world datasets in back end.

REFERENCES

[1] Abazeed A., Mamat A., Nasir M.and Ibrahim H., "Mining Association Rules from Structured XML Data", Proceedings of International Conference on Electrical Engineering and Informatics (ICEEI), 2009, pp. 376-379.

[2] Abazeed A., Mamat A., Sulaiman M.N. and Ibrahim H., "Scalable Approach for Mining Association Rules from Structured XML", Proceedings of 2nd Conference on Data Mining and Optimization (DMO), 2009, pp. 5-9.

[3] Buchner A.G., Baumgarten M., Mulvenna M.D., Bohm R. and Anand S.S., "Data Mining and XML: Current and Future Issues", Proceedings of First International Conference on Web Information Systems Engineering, 2000, pp. 131-135.

[4] Cao Y., Yang L. Yang Y., Chen H. and Liu N., "Machine Tool Distributed Cooperative Design System Based on Extended MVC-Based Web Application Framework and XML Interoperable Information Model", Proceedings of International Conference on Internet Computing in Science and Engineering (ICICSE), 2008, pp. 423-428.

[5] Cleveland F.M., "Information Exchange Modeling (IEM) and eXtensible Markup Language (XML) Technologies", IEEE Power Engineering Society Winter Meeting, 1., 2000, pp. 145-150.

[6] Czejdo B., Miller R., Taylor M. and Rusinkiewicz M., "Distributed Processing of Queries for XML Documents in an Agent Based Information Retrieval System", Proceedings of International Conference on Digital Libraries: Research and Practice, 2000, pp. 246-253.

[7] Gongxing W., "A Study on the Mining Algorithm of Fast Association Rules for the XML Data", Proceedings of International Conference on Computer Science and Information Technology, 2008, pp. 204-207.

[8] Krishnamurthy K. and Esterline A.C., "A System that Integrates Agent Services with Web Services and a Distributed Event-Based System for Processing XML Information", Proceedings of the IEEE SoutheastCon, 2010, pp. 367-370.

[9] Lee K., Min J. and Park K., "A Design and Implementation of XML-Based Mediation Framework (XMF) for Integration of Internet Information Resources", Proceedings of 35th IEEE Annual Hawaii International Conference on System Sciences (HICSS-35'02), 2002, pp. 2700-2708.

[10] Li X. Y., Yuan J. S. and Kong Y. H., "Mining Association Rules From Xml Data With Index Table", Proceedings of International Conference on Machine Learning and Cybernetics, 2007, pp. 3905-3910.

[11] Luttenberger N., Reuter F. and Koberstein J., "XML Language Binding Support for Pervasive Communication in Distributed Virtual Shared information Spaces", Proceedings of the 2nd IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMW), 2004, pp. 181-186.

[12] Maldonado J.A., Robles M. and Crespo P., "Integration of Distributed Healthcare Records: Publishing Legacy Data as XML Documents Compliant with CEN/TC251 ENV13606", 16th IEEE Symposium on Computer based Medical Systems, 2003, pp. 213-218.

[13] Memon Q.A. and Khoja S., "XML Implementation of Role Based Control in Healthcare Adhoc Networks", Proceedings of International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications, 2007, pp. 1223-1226.

[14] Moreno J.M., Fernandez D. R. and Paya A.S., "An Approach based on XML for Communication in Home Healthcare Systems", Proceedings of 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 2007, pp. 5911-5914.

[15] Pietkiewicz T., Kawalec A. and Krenc K., "An Application of the Content Description Language XML for Modeling of Information Structures of Electronic Intelligence System", Proceedings of 11th Intl. Radar Symposium (IRS), 2010, pp. 1-4

[16] Porkodi R., Bhuvaneswari V., Rajesh R. and Amudha T., "An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm", Proceedings of IEEE International Advance Computing Conference (IACC 2009), Patiala, India, 2009, pp. 1510-1514.

[17] Qi C. and Ming H., "XML-Based Data Mining Design and Implementation", Proceedings of International Conference on Computer Design and Applications (ICCDA), 2010, 4 pp. 610 613.

[18] Seals M., "The Use of XML in Healthcare Information Management", Journal of Healthcare Information Management, 14, No. 2, 2000, pp. 85-95.

[19] Shammary A.D. and Khalil I., "A New XML-Aware Compression Technique for Improving Performance of Healthcare Information Systems Over Hospital Networks", Proceedings of 32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina, August 31 - September 4, 2010, pp. 4440-4443.

[20] Song C. and Ma K., "Applications of Data Mining in the Education Resource Based on XML", Proceedings of Intl. Conf. on Advanced Computer Theory and Engineering (ICACTE), 2008, pp. 943-946.

[21] Termier A., Rousset M.C. and Sebag M., "TreeFinder: a First Step Towards XML Data Mining", Proceedings of IEEE International Conference on Data Mining ICDM, 2002, pp. 450-457.

[22] Ting C, Xiao N. and Weiping Y., "The Application of Web Data Mining Technique in Competitive Intelligence System of Enterprise Based on XML", Proceedings of Third International Symposium on Intelligent Information Technology Application, 2009, pp. 396-399.

[23] Win C. N. and Hla K. H. S., "Mining frequent patterns from XML data", Proceedings of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT), 2005, pp. 208-212.

[24] Xinliang L., Xinyuan L. and Shimin W., "XML-Based Transformation Research Between Mining Heterogeneous Data and the Relational Data", Proceedings of International Conference on Computer Application and System Modeling (ICCASM), 2010, pp. 13 292-294.

[25] Ya-qin F. and Wen-yong F., "XML in Web Data Mining Application", Proceedings of WASE International Conference on Information Engineering (ICIE), 2010, pp. 53-56.

[26] Zhang P., and Chen J., "Study and Application of Web Data Mining Based on XML", Proceedings of International Conference on Educational and Network Technology (ICENT), 2010, pp. 294-297.