Towards Data Quality and Data Mining Using Constraints in XML

Md. Sumon Shahriar and Sarawat Anam

University of South Australia, Adelaide, Australia
and Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh
E-mail: shamy022@students.unisa.edu.au, ratna_3001@yahoo.com

Abstract

Quality data is necessary for different data mining techniques and reversely, data mining techniques can be utilized to measure quality of data. Data mining and data quality issues got much attention for relational data in past. But, as a massive amount of data is being stored and represented over the web in XML, the issue of data quality for mining purposes and also using data mining techniques for quality measures get research interest. We propose two important interrelated issues: how quality XML data is useful for data mining in XML and how data mining in XML is used to measure the quality data for XML. When we address both issues, we consider XML constraints because constraints in XML can be used for quality measurement in XML data and also for finding some important patterns and association rules in XML data mining. We mainly address the theoretical framework for data quality and data mining for XML. Our research is towards the broader task of data mining and data quality for XML data integrations.

Keywords: XML, Data mining, Data quality, XML constraints

1. Introduction

Quality data is important for data mining in any data model[1]. To find some interesting patterns and rules in data mining with accuracy, semantically correct, consistent and complete data is necessary[2]. Oppositely, data mining techniques with the help of some rules and patterns[9] can be used for some quality measurement of data. In quality measurement of data, integrity constraints play a significant role. Similarly, in data mining processes, use of constraints is useful for finding patterns and rules.

In recent years, XML[13] is a widely used data representation and storage format over the web and hence the issues of data quality and the task of data mining processes are getting significant attention to the database community[10, 11, 12]. We consider how quality data in XML is necessary to data mining in XML and also how data mining in XML can be important to qualify XML data. We investigate these issues with XML constraints[7, 8]. In XML, the important integrity constraints are XML keys, XML functional dependency, XML foreign keys, XML inclusion dependencies, XML multi valued dependencies[15, 16, 17, 18, 19, 20, 21, 22, 24, 23, 25].

In some XML data quality measurements, some XML constraints are used but the definitions for XML constraints are varied. Similarly, in some XML data mining, XML constraints are used with different approach of XML constraints. We give here examples to motivate the research issues.

We first give an example how data quality(utilizing XML constraints) affects data mining in XML.

Example 1. Consider the document type definition (DTD)[1] D in Fig.1 and its conforming XML document T in Fig.2. We represent the customer information of a company using DTD D. We want to know the following information from data:

"Classify the profession of customers and offer some credit rewards to them."

We see that the structure of the DTD allows missing values of "profession" in the conforming XML documents because of the notation '?' for "profession" element. In the XML document T, there are some missing values of profession. The missing or incomplete values in the document will surely affect in finding the professions of customers and hence classification.

Now our question is how can we restrict this incomplete data. Surely, we can do it by omitting '?' from the DTD(meaning that the profession must appear in the document) or using integrity constraints.

$$
\begin{array}{l}
<!ELEMENT\ db\ (cust)+> \\
<!ELEMENT\ cust(custID, info, items+)> \\
<!ELEMENT\ info(name, addr, email, profession^{?})> \\
<!ELEMENT\ name\ (\#PCDATA)> \\
<!ELEMENT\ addr\ (\#PCDATA)> \\
<!ELEMENT\ email\ (\#PCDATA)> \\
<!ELEMENT\ profession\ (\#PCDATA)> \\
<!ELEMENT\ custID\ (\#PCDATA)> \\
<!ELEMENT\ items\ (\#PCDATA)>
\end{array}
$$

Figure 1. XML DTD D

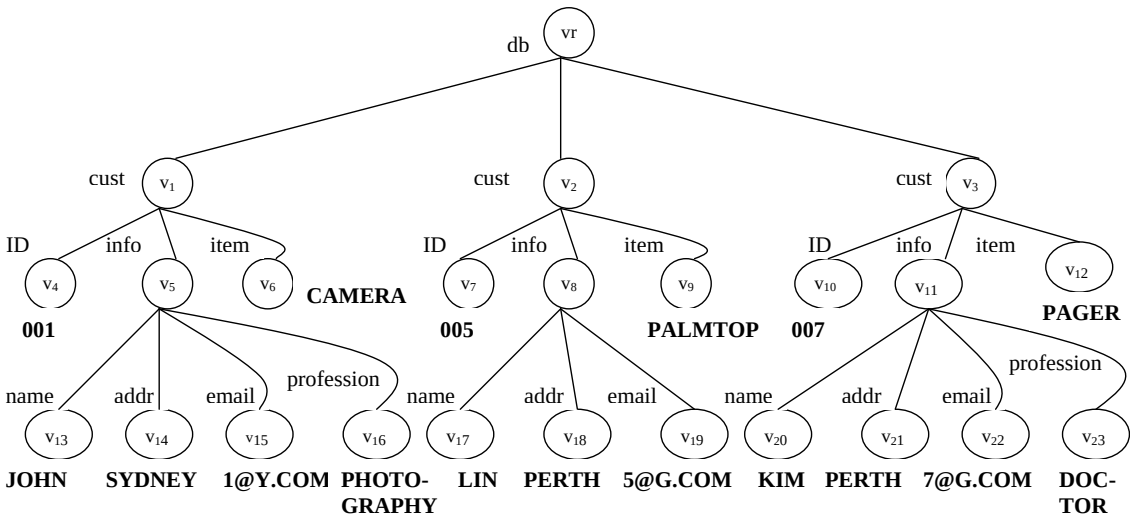

Figure 2. XML Tree T conforming to D

We can impose some constraints on the DTD. For example, we say that `custID` is the XML key and "custID determines profession" as XML functional dependency. Then we restrict that

[1]We assume that the reader is familiar with XML schema definitions such as XML DTD and XML Schema.

profession values must appear in the document according to the definition of XML functional dependency. Thus imposing constraints help to improve data quality of XML and hence help efficient data mining in XML.

Observation 1: *Quality data can be measured by XML constraints and is also needed for data mining in XML.*

Now we give another example where data mining(with the help of constraints) is used for data quality measurement in XML.

Example 2. Consider the XML document T in Fig.1 that conforms to the DTD D in Fig.2. We want to mine data for finding some rules. For example we want the following association rule:

"If the profession is photography, it is likely that he should buy a camera."

This association rule means "profession→ item" where LHS is "profession" and RHS is "item". If we observe the XML document, we see that there is a missing LHS or profession for the customer "LIN". Surely the missing or empty values affects mining the association rules. However, if we impose the constraints like "ID functionally determines profession" and profession must appear in the document as XML functional dependency, then we don't get the missing values of profession. Note that we can't say "profession functionally determines item" because this dependency is not true in the document as, for example, a photographer can buy more that one item. Thus we see how constraints in data mining can help to check quality of data.

Observation 2: *Some quality of data can be determined by data mining using XML constraints.*

Our paper is organized as follows. In next section, we give the related research issues. In section 3, we give the basic definitions. We give the proposed research framework in section 4. We give the detailed proposed methods to solve the research issues in the framework in section 5. We conclude with some remarks in the last section.

2. Related Work

Data quality is well studied topic in relational database[2, 3, 4, 5, 6]. Most research uses integrity constraints for data quality and also for data cleaning which is a task of improving data quality. In relational database, data quality mining got attention in [9, 10].

In recent years, data quality and data mining issues in semi structured data, specifically in XML, are getting much importance[11, 12]. However the utilization and characterization of XML integrity constraints for data quality and data mining purposes respectively are still of limited use. Our research is close to some previous work[2, 9, 10]. But our research is very different in the sense that we use XML integrity constraints in data quality for mining purposes and reversely data mining techniques for measuring quality of data in XML.

3. Basic Definitions

We give some basic definitions.

Definition 3.1 (Constraints,C) *By constraints, we mean keys, functional dependency, inclusion dependency and foreign key. We also mean constraints over schema and documents conform to schema and also satisfy constraints.*

Definition 3.2 (Data Quality,DQ) *By data quality, we mean the completeness and consistency of data which can be achieved by constraints, C.*

Definition 3.3 (Data Mining,DM) *We define data mining using some constraints, C. Constraints can be characterized for mining purposes.*

4. Proposed Framework

In this section, we show the proposed framework for XML.

Definition 4.1 (XML Constraint Framework,XCF) *In XML, constraint framework consists of XML keys, XML functional dependency, XML inclusion dependency, and XML foreign key. Sometimes, XML multi-valued dependency can be used.*

Definition 4.2 (XML Data Quality Framework,XDQF) *We use constraints in XCF to measure the quality of XML data. We consider completeness and consistency of XML data.*

Definition 4.3 (XML Data Mining Framework,XDMF) *By XML data mining framework, we mean the use of constraints in XCF for some mining purposes.*

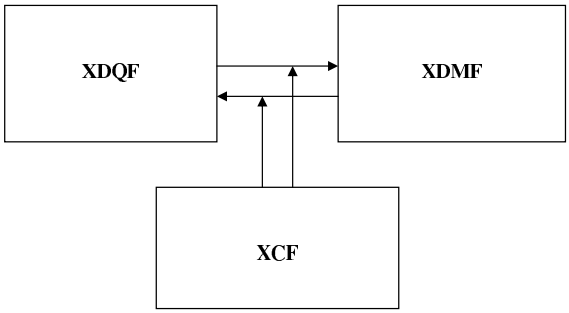

Figure 3. XML data quality and mining framework with constraints

In figure 3, we show how the XML data quality framework (XDQF)and XML data mining framework (XDMF) use the XML constraints framework (XCF). Either using XDQF for XDMF or XDMF for XDQF, XCF is utilized.

5. Proposed Methods

We now show our proposed methods of solving problems.

5.1. Quality Data for Data Mining in XML

We show how quality data is needed for data mining in Fig.4. While we discuss the quality issues in data, we use the constraints to measure those issues.

5.1.1 XML schema

We consider the XML schema for XML data. The popular XML schema definitions are XML document type definitions(DTD)[13] and XML Schema[14]. The XML documents should conform to the XML schema.

5.1.2 XML Constraints

The important XML constraints like XML keys, XML functional dependency, XML multi-valued dependency, XML inclusion dependency etc. can be used to measure quality of data in XML. When we use these constraints, we need to define them in such a way that completeness and consistency issues in XML data are captured. Also, we need to consider that the definitions for XML constraints should be over the XML schema definitions.

5.1.3 XML Data Quality Metrics for Data Mining

In this stage, we need to identify the metrics for data quality in data mining XML. The outcome of this stage will be the analytical results for measurements of XML data quality in data mining.

5.1.4 Feedback for Redefining XML schema for Data Quality Improvements

This stage recommends improvements of designing the XML schema with some constraints to improve the data quality for data mining purposes in XML.

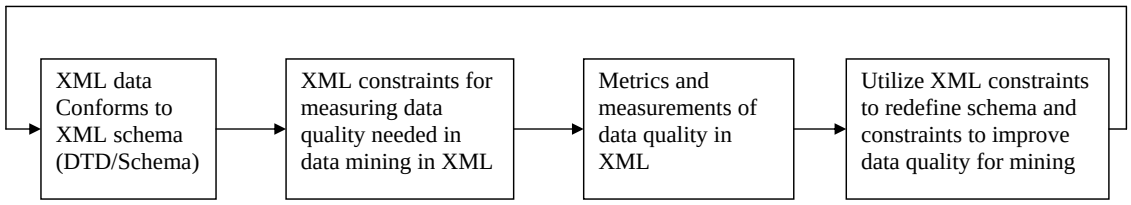

Figure 4. Processes of using quality data characterized by XML constraints for data mining in XML

5.2. Data Mining for Quality Data in XML

We discuss the processes how data mining techniques with the help of XML constraints are used for data quality measuring.

In this method, we also use XML schema and its conforming document as inputs. However the use of XML constraints for mining purposes is different in the sense that we characterize XML constraints as mining association rules.

5.2.1 XML Constraints

We use XML functional dependency, XML multi-valued dependency etc. for finding some interesting patterns and association rules in the XML documents.

5.2.2 XML Data Mining Measurements

In this stage, we measure XML data mining parameters like support and confidence values. We then make an analytical result how data mining techniques contributed towards data quality measures.

5.2.3 Feedback to Redefine XML Constraints in Mining Data in XML

We further assess how XML constraints can be redefined over XML schema to enhance the data mining features and hence discovering data quality in XML.

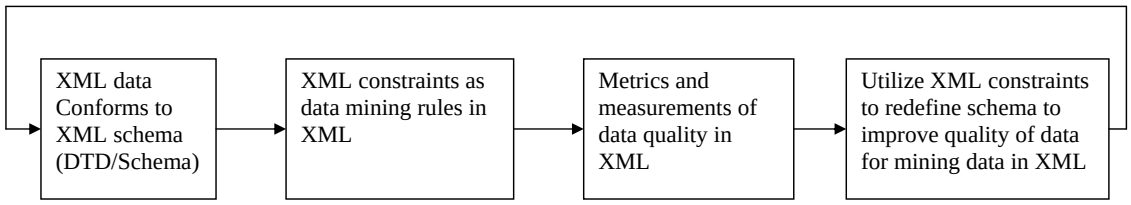

Figure 5. Processes of data mining using XML constraints for data quality in XML

5.3. Process Iterations

We discussed how data quality data affects data mining and also how data mining can be used for quality data measurements. If we observe the figures 3 and 4, we see that both methods have iterative feedback mechanism that helps to improve either data quality or data mining processes in XML. This incremental and iterative process helps data quality and also helps data mining in XML.

6. Conclusions

We propose a novel framework for data quality and data mining together in XML data model. While proposing, we consider XML constraints. The important XML constraints are characterized for both data quality and data mining processes. This paper is a theoretical framework where problems are identified, rather than solutions. We argue that the proposed framework can be implemented to solve the observations found in the introduction. Our combined framework can work for XML data integration, XML data warehousing, XML data transformation, and XML data mediation. Moreover, this approach can have significant use in data cleaning that in turn helps in data quality and data mining.

References

[1] S. Abiteboul, R. Hull, and V. Vianu, *Foundations of Databases*, Addison-Wesley, 1995.

[2] C. Gao, F. Wenfei, G. Floris, J. Xibei and M. Shuai, *Improving data quality: consistency and accuracy, Proceedings of the 33rd international conference on Very large data bases, VLDB Endowment,* Vienna, Austria, 2007.

[3] P. Bohonnon, W. Fan, F. Geerts, X. Jia, A. Kementsietsidis, *Conditional Functional Dependencies for Data Cleaning, ICDE*, 2007.

[4] W. Fan, F. Geerts, X. Jia, A. Kementsietsidis, *Conditional Functional Dependencies for Capturing Data Inconsistencies, ACM TODS*, 33(2), 2008.

[5] W. Fan, F. Geerts, X. Jia, *Semandaq: A Data Quality System Based On Conditional Functional Dependencies, VLDB*, 2007.

[6] W. Fan,*Dependencies Revisited for Improving Data Quality(invited paper), ACM PODS,* 2008.

[7] W. Fan, *XML Constraints: Specification, Analysis, and Applications, DEXA,* 2005, pp.805-809.

[8] P. Buneman, W. Fan, J. simeon and S. Weinstein,*Constraints for Semistructured Data and XML, SIGMOD Record,* 2001, pp. 47-54.

[9] D. Luebbers, U. Grimmers and M. Jarke, *Systematic Development of Data Mining-based Data Quality Tools,* VLDB, 2003.

[10] O. Hipp, U. Guntzer, and U. Grimmer,*Data Quality Mining: Making a Virtue of Necessity*, In Proc. of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), 2001.

[11] A. G. Buchner, M. D. Mulvenna, S. S. Anand, M. Baumgarten and R. Bohm, *Data Mining and XML: Current and Future Issues,* WISE'00.

[12] M.M. Khaing and N. Thein, *An Efficient Association Rule Mining For XML Data,* SICE-ICASE, 2006, pp. 5782-5786.

[13] Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen, Extensible Markup Language (XML) 1.0., World Wide Web Consortium (W3C), Feb 1998.

[14] Henry S. Thompson, David Beech, Murray Maloney, and Noah Mendelsohn, XML Schema Part 1:Structures, W3C Working Draft, April 2000.

[15] W. Fan and J. Simeon, *Integrity constraints for XML, PODS*, 2000, pp.23-34.

[16] W. Fan and L. Libkin, *On XML Integrity Constraints in the Presence of DTDs, Journal of the ACM*, 2002, vol.49, pp. 368-406.

[17] M. L. Lee, T. W. Ling, and W. L. Low, *Designing Functional Dependencies for XML, EDBT,* LNCS 2287, 2002, pp.124-141.

[18] M. Arenas and L. Libkin, *A Normal Form for XML documents, ACM PODS, 2002,* pp. 85-96.

[19] S. Hartmann and S. Link, *More Functional Dependencies for XML,ADBIS*, 2003, LNCS 2798, pp.355-369.

[20] M. Vincent and J. Liu, *Functional Dependencies for XML, APWEB,* LNCS 2642, 2003.

[21] J. Liu, M. Vincent and C. Liu, *Local XML Functional Dependencies, WIDM*, 2003, pp. 23-28.

[22] M. Vincent, J. Liu and C. Liu, *Strong Functional Dependencies and Their Application to Normal Forms in XML, ACM TODS*, 2004, pp. 445-462.

[23] J. Liu, M. Vincent and C. Liu, *Functional Dependencies, From Relational to XML, PSI,* LNCS 2890, 2003.

[24] M. W. Vincent, J. Liu and M. Mohania, *On the equivalence between FDs in XML and FDs in relations, Acta Informatica*, 2007, pp.207-247.

[25] P. Buneman, S. Davidson, W. Fan, C. Hara and W. C. Tang, *Keys for XML, WWW10*, 2001, pp.201-210.

Biography

Md.Sumon Shahriar: Sumon Shahriar is currently PhD researcher in Data and Web Engineering Lab, School of Com puter and Information Science, University of South Australia. He achieved his Bachelor of Science (Honours) and Master of Science (Research) degrees both with first class in Computer Science and Engineering from University of Dhaka, Bangladesh. His research interests include XML database, Data Integration, Data Quality and Data Mining.

Sarawat Anam: Sarawat Anam got her bachelor's degree in Computer Science and Engineering from Rajshahi University of Engineering and Technology ,Bangladesh. Her research interests include Artificial Intelligence, Database and XML, web technologies.