RESEARCHERID
THOMSON REUTERS

[Hema Priya, 6(3): March 2019]
DOI- 10.5281/zenodo.2616839

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

XML BASED APPROACH FOR NEED OF QUALITY DATA FOR DATA MINING

Hema Priya K E[*1], Prathibaa.N.S[2], Anson Sounder[3] & Dhinesh Kanna R K[4]
[1]MCA, M.Phill, [2]BSC CSA, [3]BSC CSA & [4]BSC CSA
[1]BSC Computer Science and Applications,
[1]Sri Krishna Arts And Science College, Coimbatore, India

ABSTRACT

Traditionally, quality of data plays a vital role in data mining. Since data mining is also used for measuring quality of data. So XML based approach for quality data is more useful. As large amount of data are represented in web as XML format only. Here we proposed an approach based on how the quality of data is used for data mining in XML and how data mining for XML is used to measure quality of data. XML constraints are considered as those constraints used for finding the patterns and some rules. This paper focus on theoretical framework for data mining and data quality for XML integrations.

Keywords: *Data Mining, XML, quality of data, data integration.*

I. INTRODUCTION

The term Quality of data is important for data mining in data model[1]. For finding some patterns and rules for data mining with accuracy, semantically correct, consistent and complete data is necessary[2]. Oppositely, data mining processes can be used for quality measurement of data with the help of some rules and patterns[9]. For quality measurement of data, integrity constraints is to be considered. Also in data mining processes use of constraints are useful for finding patterns and rules.

In recent years, XML is a widely used for the data representation and storage format among the web and task of data mining processes with significant attention to the database. We consider how XML quality data is necessary for data mining in XML and how XML data mining is important for qualify XML data. We investigate these issues in the XML constraints. In XML constraints were XML keys, XML functional dependency, XML inclusion dependencies, XML foreign keys, XML multi valued dependencies. There are many standards for these constraints and the research of those constraints are still exits.

In some XML data quality measurements, some XML constraints are used but again the definitions for XML constraints are varied. In some XML data mining, constraints are used with different approach of XML constraints. Here examples for the research issues is given.

Example :
Consider the XML document which contains profession details that conforms to the DTD. We want to mine data for finding some patterns.

If profession is like photography, it is likely that he should buy a camera. This association rule describes "profession→ item" where LHS is profession and RHS is item. Then we observe the XML, we see there is a missing LHS or profession for the customer LIN. Surely the missing values affects mining the association rules. However, if we impose the constraints like "ID functionally determines profession" and profession must appear in the document, 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 than one item. Thus how constraints in data mining is used to check quality of data.

ISSN 2348 – 8034
Impact Factor- 5.070

Observation :
The quality of data is determined by data mining using XML constraints of this example.

```
<!ELEMENT info(name,addr,email)>
<!ELEMENTcust(custID, inf o, items+) >
```

Figure 1. XML DTD D

II. RELATED WORK

The quality of data are the basic for the XML Data. Data quality and data mining are well studied topic in relational database. Use of integrity constraints is appeared in some research[2].The quality of data can be improved by using data mining. The data quality and data mining issues in semi structured data and in XML, are getting much importance.

Even the utilization and characterization of XML integrity constraints for data quality and data mining purposes respectively are still of limited use. But our research is that we use XML integrity constraints in data quality for mining purposes and reversely data mining techniques for measuring quality of data in XML. In this we are going to discuss the above mentioned topics like Data Quality[Quality of Data] and Data Mining.

III. BASIC DEFINITIONS

We give some basic definitions and notation which are needed for the rest of the paper.

3.1 Constraints

Here constraints means keys, functional dependency and foreign key. We also mean constraints over schema and documents conformed to schema and also satisfied by constraints. Constraints are used for the key functionality in Data.

3.2 Data Quality

Data quality means completeness and consistency of data which can be achieved by constraints. The quality of data is achieve by completeness of the constraints used for the Data Mining.

3.3 Data Mining

Usually data mining can be defined using some constraints. Constraints can be characterized as some association rules for mining purposes. The purpose of following the rules to get the Quality the Data.

IV. PROPOSED FRAMEWORK

The proposed framework for XML is explained below.

4.1 XML Constraint Framework

In XML, constraint framework consists of XML keys, XML functional dependency, XML dependency, and XML foreign key. Sometimes, XML multi-valued dependency can be used. XML constraint are basically used to attain the Data Quality.

4.2 XML Data Quality Framework

We use constraints in XML Constraint Frameworkto measure the quality of XML data. We consider completeness and consistency of XML data. We can achieve the quality of data by the completeness of data we use and the measure of data quality.

4.3 XML Data Mining Framework

XML data mining framework means the use of constraints in XML Constraint Frameworkfor some mining purposes. By XML mining the main aim is to attain Data Quality in XML.

V. PROPOSED METHODS

We now show our proposed methods of solving problems found in the observations.

5.1. Quality Data for Data Mining in XML

We show how quality data is needed for data mining in Fig.3. While we discuss the quality for issues in data, we use the constraints to measure those issues of the XML.

XML schema is considered for XML data. The most popular XML schema definition is XML document type definitions[13] and XML Schema[14]. The XML documents should match to the XML schema.

XML Constraints: The XML constraints are keys, functional dependency, multivalued dependency, inclusion dependency are 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 and ensure the data quality. Also, we need to consider that the definitions for XML constraints should be over the XML schema definitions [DTD].

Figure 2.Use of quality data characterized by XML constraints for data mining

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.

We also use XML schema and its conforming documents as inputs. Since the use of XML constraints for mining purposes is different that we characterize XML constraints as mining association rules.

XML Constraints

It uses XML functional dependency and XML multi-valued dependency.

XML Data Mining Measurements

Here we measure XML data mining parameters thatsupport and confidence values. Then we make an analytical result for how data mining techniques contributed towards data quality measures. Feedback to Redefine XML Constraints in Mining.

Data in XML

Then we assess how XML constraints are redefined over XML schema to enhance the data mining features and hence discovering data quality in XML.

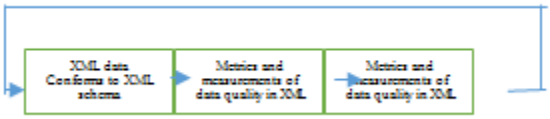

Figure 3. Data mining using XML constraints for XML data quality

5.3. Process Iterations

Here we discussed how the data quality affects data mining and also how data mining can be used for quality data measurements. Both methods have feedback mechanism which helps to improve either data quality or data mining in XML. This incremental and interactive process helps data quality and also helps data mining in XML.

VI. CONCLUSIONS

We proposed a novel framework for data quality and data mining together in XML data model. We consider XML constraints. The main XML constraints were characterized for both data quality and data mining. This paper is a 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. This combined framework can work for XML data integration, XML data warehousing, XML data transformation, and XML data mediation.

REFERENCES

1. *R. Hull and V. Vianu, Foundations of Databases, Addison-Wesley, 1995.*
2. *C. Gao, F. Wenfei, G. Floris Improving data quality- consistency and accuracy, Proceedings of the 33rd international conference on Very large data bases, VLDB Endowment, Austria, 2007.*
3. *P. Bohonnon, W. Fan, F. Geerts, Conditional Functional Dependency for Data Cleaning, ICDE, 2007.*
4. *W. Fan, F. Geerts, Conditional Functional Dependencies for Capturing Data Inconsistencies, ACM TODS, 2008.*
5. *W. Fan, F. Geerts, Semandaq: A Data Quality System Based On Conditional Functional Dependencies, VLDB, 2007.*
6. *W.Fan,Dependency Revisited in Improving Data Quality, 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.*
9. *U. Grimmers, Luebbers and M. Jarke, A systematic Development for Data Mining based Data Quality Tools, VLDB, 2003.*
10. *O. Hipp, U. Guntzer, and U. Grimmer,Data Quality Mining- Making Virtue of Necessity, In 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), 2001.*
11. *S. S. Anand, M. Baumgarten, Data Mining and XML: Current and Future Issues.*
12. *M.M. Khaing and N. Thein, An Efficient Association Rule Mining For XML Data, SICE-ICASE, 2006.*
13. *Tim Bray, Jean Paoli, Extensible Markup Language (XML) 1.0., World Wide Web Consortium (W3C), Feb 1998.*
14. *Henry S. Thompson, Murray Maloney, and Noah Mendelsohn, XML Schema- Structures, W3C Working Draft, April 2000.*
15. *W. Fan and J. Simeon, Integrity constraints for XML, PODS, 2000.*
16. *W. Fan, On XML Integrity Constraints in the Presence of DTDs, Journal of the ACM, 2002, vol.49.*
17. *M. L. Lee, T. W. Ling,, Design Functional Dependency for XML, EDBT, LNCS 2287, 2002.*
18. *M. Arenas, A Normal Form for XML documents, ACM PODS, 2002.*
19. *S. Hartmann and S. Link, More Functional Dependencies for XML,ADBIS, 2003.*
20. *J. Liu, Functional Dependency for XML, APWEB, 2003.*
21. *J. Liu and C. Liu, Local XML Functional Dependency, WIDM, 2003.*
22. *J. Liu and C. Liu, Strong Functional Dependency and Application to Normal Forms in XML, ACM TODS, 2004.*
23. *J. Liu and C. Liu, Functional Dependency, From Relational to XML, PSI, 2003.*
24. *J. Liu and M. Mohania, On equivalence between FDs in XML and FDs in relations, ActaInformatica.*
25. *P. Buneman, W. Fan, C. Hara and W. C. Tang, Keys for XML.*