Data Cleaning and XML : The DBLP Experience

Wai Lup Low, Wee Hyong Tok, Mong Li Lee, Tok Wang Ling
School of Computing
National University of Singapore
{lowwailu,tokweehy,leeml,lingtw}@comp.nus.edu.sg

With the increasing popularity of data-centric XML, data warehousing and mining applications are being developed for the rapidly burgeoning XML data repositories. Data quality will no doubt be a critical factor for the success of such applications. Data cleaning, which refers to the processes used to improve data quality, has been well researched in the context of traditional databases. We developed a knowledge-based framework for data cleaning relational databases in [1]. In this work, we present a novel attempt to apply this framework to XML databases. Our experimental dataset is the DBLP database, a popular online XML bibliography database used by many researchers.

Cleaning XML databases poses new challenges and problems not faced in cleaning relational databases. The first challenge is to map the complex structures in XML to fact templates used in expert systems. We develop an algorithm to map the DBLP DTD specification to a collection of corresponding fact templates. The algorithm is generic enough to accommodate other DTDs with minor or no modifications.

Another challenge is to determine if XML parsing using standard APIs is more efficient than simple text parsing. We compare the performance of parsing simple XML structures using the SAX API with text parsing methods. Experimental results show that if the structure of the XML database is simple and static, parsing the XML database with regular expression matching methods may offer large performance gains.

Since the knowledge-based framework makes use of a sliding window of sorted data to detect anomalies within the neighbourhood, the ability to efficiently feed the ordered data into the knowledge base is very important. The next challenge is to find efficient indexing techniques to achieve this. Replicating data in the index will reduce the time taken to clean the XML database. However, data replication may bring about data consistency problems. An associated challenge is to quantify the gain in performance and the increase in storage requirements due to the data replication. We study the performance and storage requirements of three set-ups for the index. We explain why using a RDBMS for the index offers the best in terms of scalability and maintainability, though it may not be the fastest.

We used the proposed approach to clean the DBLP XML database of inexact duplicate author names (which has been identified as a never-ending problem). The sampled precision was about 40%, but this figure does not include cases where even humans cannot ascertain if they are true duplicates (which can only raise the precision). Better recall and precision can be achieved with the injection of more domain knowledge as shown in [1]. We also illustrate how the approach can efficiently verify domain-specific business rules.

The contributions of this work are:

1. We present a novel approach to improve XML data quality by using a knowledge-based system to clean the data.

2. Review the necessity of using of XML parsers for simple XML structures and compare their performance with parsing via regular expression matching.

3. Examine different ways to map DTD specifications to fact templates in knowledge bases.

4. Investigate the performance of various ways to index the data in XML documents, and the feasibility of replicating the data in the index for faster processing.

References

[1] W. L. Low, M. L. Lee, and T. W. Ling. A Knowledge-Based Approach for Duplicate Elimination in Data Cleaning. *Information Systems : An International Journal. Special Issue on Data Extraction, Cleaning, and Reconciliation.*, 2001.