A Methodology for Building XML Data Warehouses

Laura Irina Rusu, La Trobe University, Australia
J. Wenny Rahayu, La Trobe University, Australia
David Taniar, Monash University, Australia

ABSTRACT

Developing a data warehouse for XML documents involves two major processes: one of creating it, by processing XML raw documents into a specified data warehouse repository; and the other of querying it, by applying techniques to better answer users’ queries. This paper focuses on the first part; that is identifying a systematic approach for building a data warehouse of XML documents, specifically for transferring data from an underlying XML database into a defined XML data warehouse. The proposed methodology on building XML data warehouses covers processes including data cleaning and integration, summarization, intermediate XML documents, and updating/linking existing documents and creating fact tables. In this paper, we also present a case study on how to put this methodology into practice. We utilise the XQuery technology in all of the above processes.

Keywords: data warehouse; star schema; XML; XML schema

INTRODUCTION

In the last few years, building a data warehouse for XML documents has become a very important issue, when considering the continual growth of representing different kinds of data as XML documents (Widom, 1999; World Wide Web Consortium). This is one of the reasons why researchers became interested in studying ways to optimise processing of XML documents and to obtain a better data warehouse to store optimised information for future reference.

Many papers have analysed how to design a better data warehouse for XML data from different points of view (e.g., Widom, 1999; Goffarelli, Maio & Rizzi, 1998; Vrdoljak, Banek & Rizzi, 2003; Zhang, Ling, Bruckner & Tjoa, 2003) and many other papers have focused on querying XML data warehouse or XML documents (e.g., Fernandez, Simeon & Wadler, 1999; Deutch, Fernandez, Florescu, Levy
& Suciu, 1999), but almost all of them have considered only the design and representations issues of XML data warehouse or how to query them and very few have considered optimisation of data quality in their research.

In this paper, we propose a practical methodology for building XML documents data warehouses. We ensure that the data warehouse is one where the occurrences of dirty data, errors, duplications or inconsistencies are minimized as much as possible and a good summarisation exists. The steps cover two stages: (A) data cleaning and (B) data summarization, creating fact documents and linking all documents to create data warehouses. We use XQuery in all of the above processes. The main purpose of this paper is to show systematic steps to building an XML data warehouse as opposed to developing a model for designing a data warehouse. However, it is important to note that our proposed steps for building an XML data warehouse are generic enough to be applied to different XML data warehouse models.

The rest of this paper is organised as follows: After discussing related work, we present our proposed methodology for building XML data warehouses, then a case study exemplifies our methodology and the final section gives the conclusions.

RELATED WORK

There is a large amount of work in the data warehouse field. Many researchers have studied how to construct a data warehouse, first for relational databases (Goffarelli et al., 1998; Galhardas, Florescu, Shasha & Simon, 2000; Roddick et al., 1999; Song, Rowen, Medsker & Ewen, 2001) but in the last few years, for XML documents (Vrdoljak et al., 2003; Zhang et al., 2003), considering the spread of use for this kind of documents in a vast range of activities. Furthermore, if we think of steps in our proposed methodology, there were few attempts to solve the problem of data cleaning automation, too, but most researchers concentrated on databases field analysis.

A concrete methodology on how to construct an XML data warehouse analysing frequent patterns in user historical queries is provided in Zhang et al. (2003). The authors start from determining which data sources are more frequently accessed by the users, transform those queries in Query Path Transactions and, after applying a rule mining technique, calculate the Frequent Query Paths which stay at the base of building data warehouse schema. It was also mentioned that the final step in building a data warehouse would be to acquire clean and consistent data to feed to the data warehouse. However, there is not enough detail on how to ensure this. Although it seems to be a simple thing to do in the whole process, this is the place where corrupted or inconsistent data can slip into the data warehouse.

Another approach is proposed in Vrdoljak et al. (2003), where an XML data warehouse is designed from XML schemas, proposing a semi-automated process. After pre-processing an XML schema, creating and transforming a schema graph, the designer chooses facts for the data warehouse and, for each fact, follows a few steps in order to obtain star-schema: building the dependency graph from schema graph, rearranging the dependency graph, defining dimensions and measures and creating logical schema. In this approach, XQuery is used to query XML documents in three different situations: (i) examination of convergence and shared
A Brief Study of Approaches to Text Feature Selection
www.igi-global.com/chapter/a-brief-study-of-approaches-to-text-feature-selection/185985?camid=4v1a

Early Works in Spatio-Temporal Mining
Wynne Hsu, Mong Li Lee and Junmei Wang (2008). Temporal and Spatio-Temporal Data Mining (pp. 112-129).
www.igi-global.com/chapter/early-works-spatio-temporal-mining/30264?camid=4v1a