A THEROTICAL STUDY ON DATA EXTRACTION FOR BUSINESS INTELLIGENCE

Dhanunjaya Rao Bachalakuri

Abstract- **Though, data warehousing development is directed by procedural and architectural approaches intended at facilitating the delivery of successful results within project-defined boundaries. The World Wide Web offers public domain information which can be retrieved for instance from Web sites or online shops. The extraction from semi-structured information sources is typically done manually and is therefore very time consuming. This paper refer to how public information can be extracted automatically from Web sites, trans- formed into structured data formats, and used for data analysis in Business Intelligence systems.**

Index Terms— **Data extraction, Bussiness Intelligence, Data warehouse, XML**

I. INTRODUCTION

Advances in the information technology1 domain have guided the evolution of data processing systems from the early stages of single support and stand-alone applications to the comprehensive Business Intelligence and analytical systems of today's informational environment. Within this broad context, data warehousing defines an extensive blend of technologies emerged in early 1990s as result of advances in the area of data processing capabilities achieved in computer-based information systems. Data warehousing represents a component of the overall Business Intelligence framework, which covers an ample range of applications and tools that help analyze large amounts of data and transform it into understandable information and business knowledge. It is designed to handle an informational environment in which a series of components enable the gathering and integration of data from across the enterprise in order to provide business users with consolidated, structured data and improve the decision-making processes. The repository of this comprehensive data warehousing technology, defining the storage environment, is known as the data warehouse. The data warehouse represents a model of enterprise data, especially structured for facilitating querying and analysis processes on integrated and consolidated data. It is an essential and dominant component of data-driven decision support systems and its main goal is to enable business users to make effective tactical and strategic decisions based on factual data, by answering business questions timely and accurately. For achieving its goal, the data warehouse is defined by particular data models that specify the structure of data in the repository. These data models, optimized for querying and analysis, are created in a stable, consistent and predictable manner through different data modeling techniques. Querying and analysis are enabled by means of various types of metadata meant to describe the structure of an organisation's use of information and to attach semantics to the business processes and the resulting data. Considering its complexity level, data warehousing solutions development demands a structured and planned approach, defined in the form of a methodology, as well as an appropriate architectural framework. Methodologies are designed to achieve predictable results according to well-define requirements and to provide repeatable, trainable and consistent development processes. Architectures represent the structures that bring all the components of the data warehouse together and provide a solid basis for enterprise-wide data integration. The selection of a suitable methodology and architecture determine the overall success of the data warehouse solution implementation. Another essential aspect in the data warehouse development regards the employment of a framework able to provide a set of guidelines for describing its components and their interoperability and to support the existence of an environment of reusable processes, integration, consistency and flexibility in information delivery.

II. PROPOSED MODEL

A. Business Intelligence Technology

Main goal is to place the data warehousing theme in the overall Business Intelligence framework. For this reason, we outline briefly the historical use of data, examine the evolution of intelligent decision support systems and discuss their benefits and importance in the decision-making process. We also analyze the various definitions, architectures and development lifecycles of the Business Intelligence systems, the relation between the Business Intelligence and the data warehousing technologies, and the role of the data warehouse concept in the analytical framework. Advances in the information technology field have guided the evolution of data processing systems from the early days of single support and stand-alone applications, such as Management Support Systems, to the

comprehensive Business Intelligence and analytica technologies of today's informational environment.The Business Intelligence process covers three main process steps: data integration, data storage and data usage (see fig. 1).

Data integration covers methods to extract data from internal or external data sources such as ERP systems or database systems. The data is transferred into a processing area allowing further data transformations like data "cleaning" and data normalization. A load process contains a scheduler which regularly (e.g. daily, weekly, or monthly) uploads the processed data into the final data base storage, the data warehouse.

Data storage in a data warehouse: the basic idea of a data warehouse is to store the relevant data for decision makers in a dedicated, homogeneous database. An important characteristic of the data warehouse is the integration of heterogeneous, distributed, internal and external data. This covers the physical storage of data in a single, centralized data pool, and it also covers the subject-oriented clustering of data organized by business processes, such as sales, production, or finance. The subject oriented organization of data is called a data mart.

Data usage: to support decision making, data in a data warehouse has to be well-organized to fulfill different end-user requirements: predefined reporting for occasional users, ad-hoc data analysis for knowledge workers, or data mining for data analysts.

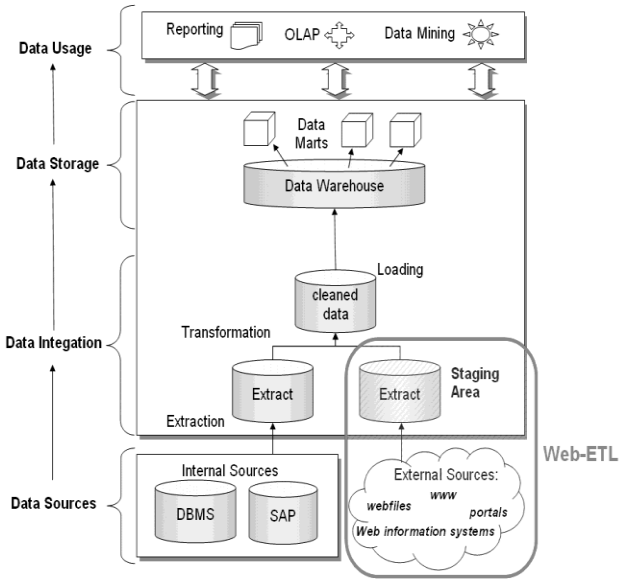

Figure 1: The Business Intelligence reference process.

At first, so-called wrappers are generated. Dynamically and independently, these "intel- ligent"3 software agents extract and translate all relevant information from HTML Web pages to a structured XML format that can be queried and processed by other programs. With Lixto, wrappers are generated in a graphical user interface with a few mouse clicks. Thus, no special programming knowledge is needed, and wrappers can be generated by non-technical personnel. Wrapper agents are typically generated by employees with the relevant business expertise for the project, e.g. from a company's marketing department.

In a second step, XML data generated by wrappers is processed in the Lixto Transformation Server [GH01], the run-time environment for Lixto wrapper agents. A wrapper in the Transformation Server retrieves the Web data automatically, with no developer interaction, based on events. Events are for example a Web page content change, or a defined schedule, such as Web data retrieval every hour or every 5 minutes. For example, if a wrapper cannot extract data from a specific Web site because the Web server is down, the wrapper generates an error message for the administrator.

Finally, the Transformation Server delivers the extracted, aggregated information into the desired formats to other Business Intelligence systems such as SAP Business Information Warehouse or Microsoft Analysis Server. Also, the Transformation Server interactively communicates with these systems using various interfaces, such as special database for- mats, XML messaging, and Web services.

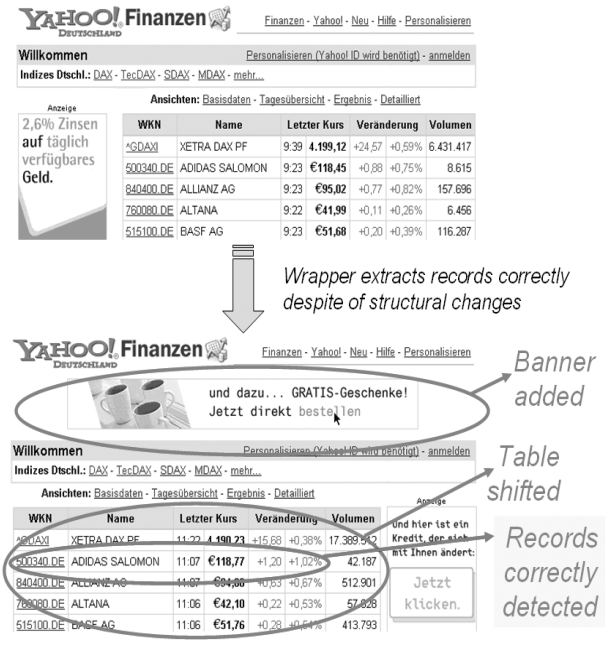

Figure 2: Wrapper robustness.

Fig. 2 shows an example: the share prices from the companies listed in the German share index DAX are to be extracted from the Web site finance.yahoo.de. After a wrapper agent was successfully generated, the layout of the Web site changed after some weeks: the table with the quoted stocks moved from left to right, and additional banners were added. For the wrapper of fig. 2, such conditions could be, the relevant area should contain the - symbol in each line" or "some specified company's names should occur". On the overview page shown in fig.3 only

the first two lines of the model description are displayed. For each model name a linked sub-page with the whole description text exists. Furthermore there is a “next”-link (“weiter”) leading to the next article overview page. The Lixto Software allows to record navigation sequences in a kind of macro recorder. During wrapper generation, only one sub-page needs to be accessed as an example and the “next”-link needs to be followed only once. The system then recognizes the similarly structured Web pages and extracts all complete model descriptions from all overview pages. The results are transformed to structured XML.

Lixto wrapper agents are embedded in the runtime-environment of the Lixto Transformation Server. This server allows post processing the XML data generated by wrapper agents. Here data from different wrappers can be aggregated, re-formatted, transformed and delivered.

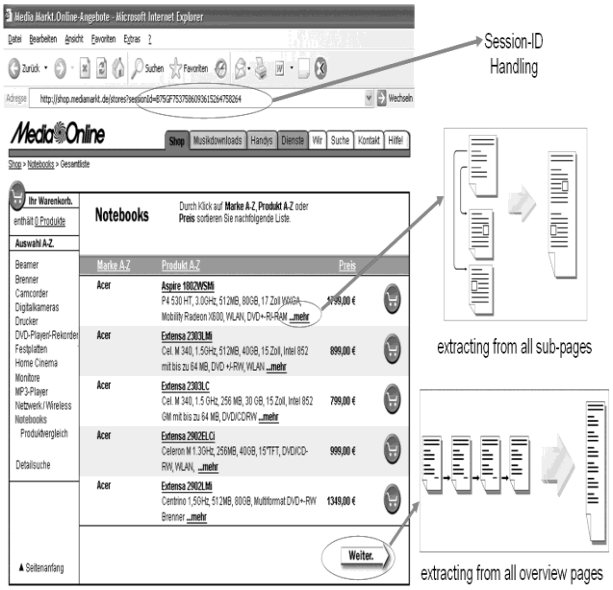

Figure 3: Extraction of all article data.

The whole process of modeling the workflow and dataflow is done in a graphical user interface in the Lixto Transformation Server. Graphical objects symbolize components, such as an integrator for the aggregation of data or a deliverer for the transmission of information to other software applications.

As an example, fig. 4 shows how Web page data is selected for extraction using the Lixto software. In the lower window of fig.4,a Web page from shop.mediamarkt.de has already been loaded in a Web browser, and the relevant information has been marked with two mouse clicks. The upper windows shows already defined logical patterns, such as article and price, arranged in a hierarchical structure. This structure corresponds to the XML output that will later be generated by the wrapper. After loading a Web page with relevant data into the Lixto software, at first a pattern named article is defined. This pattern later recognizes lines with article information. Within this line, other patterns are created, identifying information such as article manufacturer and article price.

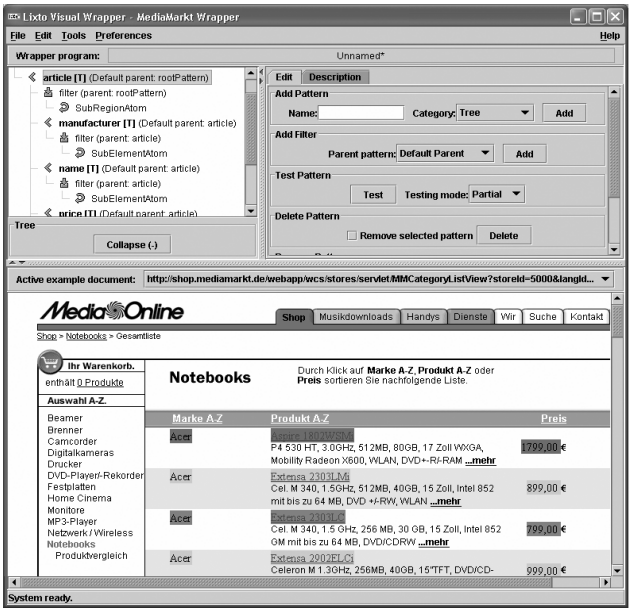

Figure 4: Visual wrapper generation with Lixto.

III. CONCLUSON

In this paper we showed how data can be automatically extracted from semi-structured web sites to obtain competitor information for decision support. We have introduced the architecture of the Lixto software for processing web data in an efficient manner. Data analysts are able to obtain knowledge about the market situation in nearly real-time. This leads to better pricing decisions, a better positioning of the company and its products on the market, and a faster reaction to competitive activities, such as product innovations, price dumping, or promotions.

REFERENCES

[1] K. Aberer, P. Fankhauser, G. Huck, and E. Neuhold. JEDI: Extracting and Synthesizing Information from the Web. In Proc. of COOPIS, pages 32–43, 1998.

[2] SAP AG. How to send XML Data to BW, ASAP for BW Acceleration. http://www.sdn.sap.com/documents/a1-8-4/HowtoSendXMLDatatoBW.pdf [accessed 2004-09-28].

[3] F.Azavant and A. Sahuguet. Building light-weight wrappers for legacy Web data- sources using W4F. In Proc. of VLDB, pages 738–741, 1999.

[4] R. Baumgartner, S. Flesca, and G. Gottlob. Visual web information extraction with Lixto. In Proc. of VLDB, pages 119–128, 2001.

[5] P. A. Bernstein, J. Madhavan, and E. Rahm. Generic Schema Matching with Cupid. The VLDB Journal, pages 49–58, 2001.

[6] Gottlob and M. Herzog. Infopipes: A Flexible Framework for M-Commerce Ap- plications. In Proc. of TES workshop at VLDB, pages 175–186, 2001.
[7] G. Gottlob and C. Koch. Monadic datalog and the expressive power of languages for Web Information Extraction. In Proc. of PODS, pages 17–28, 2002. Full version: Journal of ACM 51 (1), pages 74–113, 2004.
[8] M. Hahne, L. Burow, and T. Elvers. XML-Datenimport in das Information Ware- house bei Bayer Material Science. In Schelp, Winter, Robert (Hrsg.). Auf dem Weg zur Integration Factory, pages 231–251, 2004.
[9] R. Himmeroder, G. Lausen, B. Ludascher, and W. May. A Unified Framework for Wrapping, Mediating and Restructuring Information from the Web. In WWWCM. Sprg. LNCS 1727, pages 307–320, 1999.
[10] I.M. Nagy, Automation prototype for the development of data warehousing data structures, accepted for publishing in Procedia Technology Journal, Elsevier Publishing Ltd., ISSN: 2212-0173 (indexed ISI).Pages.

BIODATA
AUTHOR 1

Dhanunjaya Rao Bachalakuri completed his B.Tech in CSE from Rammappa Engineering College, Warangal in 2004 and M.Tech CSE from Hi-Tech Engineering College, Hyderabad in 2014, Telangana, India. His research areas of interest are Datamining, Software Engineering, Data warehousing.