Enhanced Architecture of a Web Warehouse based on Quality Evaluation Framework to Incorporate Quality Aspects in Web Warehouse Creation

Umm-e-Mariya Shah
Computer Science Department
COMSATS Institute of Information Technology Islamabad, Pakistan

Azra Shamim
Computer Science Department
COMSATS Institute of Information Technology, Islamabad, Pakistan

Madiha Kazmi
Computer Science Department
COMSATS Institute of Information Technology Islamabad, Pakistan

***Abstract*—In the recent years, it has been observed that World Wide Web (www) became a vast source of information explosion about all areas of interest. Relevant information retrieval is difficult from the web space as there is no universal configuration and organization of the web data. Taking the advantage of data warehouse functionality and integrating it with the web to retrieve relevant data is the core concept of web warehouse. It is a repository that store relevant web data for business decision making. The basic function of web warehouse is to collect and store the information for analysis of users. The quality of web warehouse data affects a lot on data analysis. To enhance the quality of decision making different quality dimensions must be incorporated in web warehouse architecture. In this paper enhanced web warehouse architecture is proposed and discussed. The enhancement in the existing architecture is based on the quality evaluation framework. The enhanced architecture adds three layers in existing architecture to insure quality at various phases of web warehouse system creation. The source assessment, query evaluation and data quality layers enhance the quality of data store in web warehouse.**

Keywords-component; Data Warehouse; Web Warehouse; Quality Assessment, Quality Evaluation Framework; Enhanced Web Warehouse Architecture; WWW

I. INTRODUCTION

Due to tremendous advances and achievements in information technology data is being generated at tremendous speed. World Wide Web plays a vital role in information retrieval. World Wide Web has grown to be a universal source and is globally used by individuals and business organizations for information sharing and exchange. Massive amount of data available on the web that is distributed, heterogeneous and semi-structured in nature. Relevant information retrieval is difficult from the web space as there is no universal configuration and organization of the web data [1], [2]. Proper management and retrieval mechanism is required to analyze the information. The distributed and heterogeneous nature of the web data tends to adapt the approach of web warehousing. Currently browsers and search engines are used for information retrieval. Due to lack of knowledge; many search engines may not fully utilize link information [1]. As a result search engines are not able to support such queries or fail to return link information [3]. Web servers can not keep track of the diverse behavior of the client's requests and does not offer the services of web personalization [1]. Therefore an intermediate storage area between the web servers and the clients proves to be a valuable resource [4]. The intermediate repository may not only serve as the storage area but also keeps track of the client's activities and helps in the web personalization [5]. Web warehousing can overcome this problem.

In this research work; the authors enhanced the basic architecture of the Web Warehouse presented in [6] on the basis of the quality evaluation framework discussed in [1]. The authors contribute towards the addition of the quality assessment layer at the time of source selection. A query evaluation layer is embedded with the query processor that may facilitate in query processing. Moreover data quality assessment layer is incorporated between the merge and the load process of the web warehousing system for maintaining high quality of data in web warehouse.

The rest of paper is organized into different sections. Section 2 provides the Literature Review regarding Data Warehouse, Web Warehouse, quality Evaluation Framework. Section 3 consists of the enhanced architecture of the web warehouse based on the quality evaluation framework discussed in section 2. Section 4 presents concluding marks.

II. LITERATERA REVIEW

A. Data Warehouse

Data Warehouse is a central repository that supports executive decision making. According to Hoffer et. al. "A data warehouse (DWH) is an 'informational database' that is maintained separately from an organization's operational database" [1], [7]. "A collection of corporate information, derived directly from operational systems and some external data sources. Its specific purpose is to support business decisions, not business operations" [1], [8]. According to Inmon, "A Data Warehouse is a subject-oriented, integrated, time-variant, non volatile collection of data in support of management decisions" [1], [9], [10], [11], [12]. Different architectures of data warehouse are discussed in [7], [8], [13], [14].

B. Web Warehouse

A Web warehouse is a combination of data warehousing technology and the web technology. According to Mattison, "It is an approach to the building of computer systems which has as its primary functions the identification, cataloguing, retrieval, (possibly) storage, and analysis of information (in the form of data, text, graphics, images, sounds, videos, and other multimedia objects) through the use of Web technology, to help individuals find the information they are looking for and analyse it effectively"[1], [16].

Web warehouse is an architecture comprising of some tools and processes necessary to build up an efficient data warehouse that works on the web and is based on web technologies. Its main functionality is the organization and analysis of stored data and its proper administration. The sources of a web warehouse are the web sites. The web warehouse stores organize and manage the information from web sources and works passively on it. The basic functions of web warehouse include Information Sharing and Intelligent Caching [4]. The contributors of Web Warehouse are Web Technology and Data Warehusing [17]. Different architecture of web warehouse is discussed in [18], [19],[20], [21].

C. Quality Evaluation Framework for a Web Warehousing System

Quality is one of the important factors for the success and survival of any system. To improve quality of a Web Warehouse, Maria et. al. proposed a quality evaluation framework in [1]. This frame work is based on [23], [24],[25] and [26].

Proper evaluation/validation of each phase of web warehouse against certain attributes to measure the quality of the system can be achieved through this framework. It shows categories and dimensions of the quality factors. Further more relevancy of these categories to the phases and sub-phases of a Web Warehouse is also discussed. Quality attributes of discussed in the framework are accessibility, interpretability, usefulness, believability, navigation, efficiency, authority, currency, availability, information-to-noise ratio, popularity, cohesiveness, integrity, reliability, functionality, efficiency, and maintainability.

III. Proposed architecture

Enhanced architecture based on the quality evaluation to improve the quality of Web Warehouse System is presented in layered approach in figure 1. The main components of the proposed architecture are described as follow:

A. Web Information Sources

The web warehouse is constructed over a heterogeneous, distributed and semi-structured web space. The data is gathered from different web sites. It undergoes the process of transformation and integration and stored in a web warehouse. The web information sources are defined at the time of making web warehouse design specification. However new web information sources can be included. It involves creation of data source view in a web warehouse along with the creation of the respective view manager. The newly added data source is then connected with the respective monitor and wrapper.

B. Source Evaluation Layer

This layer assesses the origin of the data that is selected for data extraction. Assessment is on the basis of some quality dimensions i.e. source currency, relevancy, availability, information-to-noise ratio, authority, popularity and cohesiveness. It measures the number of broken links on a web page, proportion of the useful information, prestige of the data source, relevancy of the major topics in a web page and the number of citations by other web pages. The evaluation ensures that there exist up-to-date contents in the selected source that are compatible to the user's query. The source evaluation layer then filters out the data belonging to those sources only that satisfies the evaluation criteria.

C. Monitor

This component is connected to the underlying information sources. Each data source has its monitor. It polls the web information sources periodically to detect any changes arise in them. Polling is done by comparing the snapshots and obtaining the base data changes. It collects the changes and notifies the modifications to the integrator component.

D. Wrapper

This component deals with the data extraction. It accepts the query from query processor and mines results from the underlying information sources. The result is then transformed into a specified format of a web warehousing system.

E. Integrator

This component maintains consistency between the web warehousing system and the underlying information resources. Any updated information is sent to the integrator by the monitor. Integrator integrates the information and sends the modifications to the respective view manager.

F. View Manager

It keeps the consistency between the views in a web warehouse and views of the data sources. Every view in a web warehouse has its own view manager to perform all necessary actions. Whenever a change is detected in the underlying information resource it is transferred by the monitor to the integrator. Integrator then sends the modification to the relevant view manager. The view manager then updates the relevant web view.

G. Query Processor and Evaluator

Whenever a query is initiated by a user it is transferred to the query processor and evaluator. Query evaluator assesses the query on the basis of certain quality dimensions like semantics, performance, time behavior and optimization. It ensures that the system must timely respond to a directed task and makes certain the clarity of its meaning, structure and language rules. Query processor then executes the refined query. It translates the query from the high level language to the low level language. It consults the meta data repository and directs the query to the appropriate data source.

H. Merge Process

View managers when perform actions the results are passed to the merge process. The merge process combines and sorts these results according to the user's query sequence.

I. Data Quality Assessment Layer

This layer evaluates the data in terms of certain quality dimensions. Extracted data may be erroneous and contains some inconsistencies.

The data quality assessment layer ensures that the data must be correct, comprehensive precise and stable. Thus high quality data will become available to load in the web warehouse that leads to a quality decision. This layer performs some quality control techniques as described below:

- Data Auditing and Standardization - Typically, the data in data stores and databases is inconsistent and lacks conformity. Data auditing ensures the precision and accuracy of data at the source [27]. It evaluates the data (in the source database) against a set of business rules to perform validation checks. It provides frequencies of data fields and identifies the outliers and the range of value for each attribute. The business and cleansing rules are identified in the data auditing process. The business rules may be determined by using data mining techniques which are used to uncover the patterns in the data. Outlier data is then modified as required.

- Data Linking and Consolidation - The data coming from multiple sources may be inconsistent and redundant. Data linking identifies the records that represent the same values of an entity and links them. In the consolidation process elements of matching records are combined into a complete record [27].

- Information Stewardship - The validity of information can be obtained if automated routines and business rules are implemented but they do not help for information accuracy. People and experts are needed to assure the accuracy [28]. Stewardship is "the willingness to be accountable for the well-being of the larger organization by operating in service of, rather than in control of those around us" [29]. Information stewardship is "the willingness to be accountable for a set of business information for the well-being of the larger organization by operating in service, rather than in control of those around us" [29].

- Data Cleaning - "Data cleansing is a process of identifying and removing errors or inconsistencies from the data in order to improve the quality" [30]. Data quality problems exist in each case whether data has single or multiple sources.

J. Load Process

The load process loads the high quality data to the web warehouse. This data is the result of the user's query response.

K. Web Warehouse

It is the final destination where the results are stored. Web warehouse has following components.

- Web Marts - These are designed separately for a particular department. Web marts are the subsets of a web warehouse and contains the data satisfying that particular department needs. In this way web marts help to decrease the query response time of the end user.

- Meta Data Repository - Meta data means the data about data. It is a storehouse where all the necessary information regarding extracted data is stored. It is consulted by various processes in performing their tasks. Like wrappers consult the meta data for the relevant data sources before starting the extraction process. Query processor uses the meta data repository to find the appropriate data source for the query execution phenomenon.

- Meta Data Manager - It supports the maintenance of the meta data repository. The data management and manipulation of meta data repository is handled by it.

- Web Manager - Web sites that are selected as the information sources for a web warehouse are managed by the web manager. It makes decision for the addition and deletion of the data sources. It also monitors the performance of the view managers.

L. Presentation Layer

This layer provides interface to the user. It interacts with the web warehouse and extracts the information. The data that is extracted from a web warehouse is analyzed via various tools and helps in decision making.

IV. CONCLUSION

During the web warehouse creation phase the output of one phase becomes the input of the next phase, so quality assessment is most important at various stages of the web warehouse to get a successful web warehousing system. Keeping in view the quality factor an enhanced architecture of web warehouse is proposed and discussed in this paper. The enhanced architecture increases the quality of web warehouse system by introducing source assessment, query evaluation and data quality layer. The basic architecture of a web warehouse is enhanced by embedding the source assessment layer, query evaluation layer and data quality layer. Data quality layer ensures the quality of data before loading it into warehouse. Source assessment layer is responsible for checking validity, relevancy and other quality attributes of web sources. Query evaluation layer facilitate in query processing.

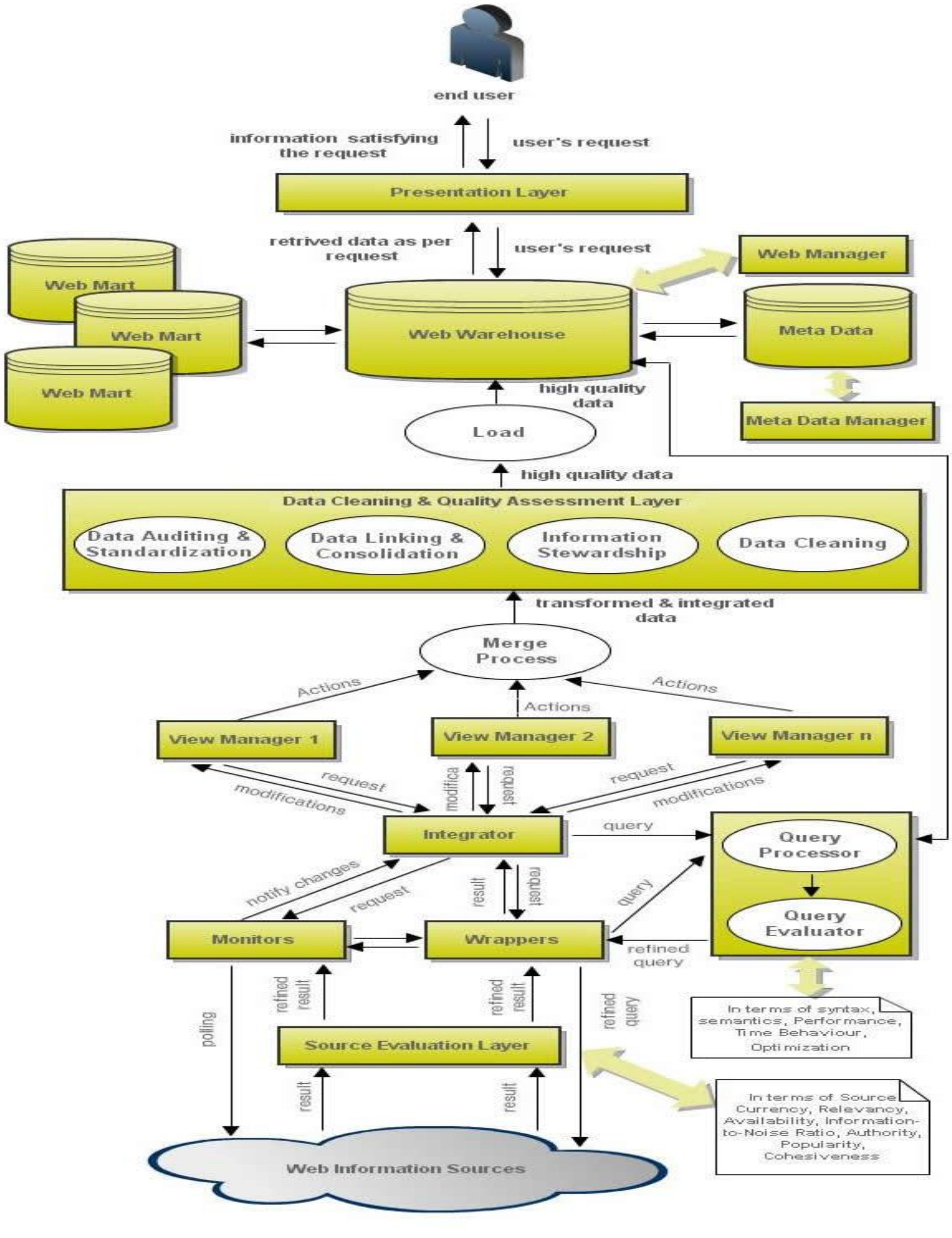

Fig. 1 Enhanced Architecture of a Web Warehouse

REFERENCES

[1] Umm-e-Mariya Shah, Maqbool Uddin Shaikh, Azra Shamim , Yasir Mehmood, Proposed Quality Evaluation Framework to Incorporate Quality Aspects in Web Warehouse Creation, Journal of Computing, Volume 3, Issue 4, May 2011.

[2] J. Dyche, “e-Data: turning data into information with data warehousing,” Addison-Wesley, Reading, MA, 2000

[3] Sourav S. Bhowmick, Wee-Keong Ng, and Ee-Peng Lim, “Information Coupling in Web Database,” Springer-Verlag Berlin Heidelberg, pp. 92-106, 1998.

[4] Kai Cheng, Yahiko Kambayashi, Seok Tae Lee and Mukesh Mohania, "Functions of a Web Warehouse," International Conference on IEEE Digital Libraries: Research and Practice, Kyoto, 2000.

[5] Masahiro Hori, Goh Kondoh, Kohichi Ono, Shin ichi Hirose, and Sandeep Singhal, Annotation-based Web Content Transcoding, http://www9.org/w9cdrom/index.html, Accessed Date: 05 Dec, 2009.

[6] Saif ur Rehman, Maqbool Uddin Shaikh, " "Web Warehouse: Towards Efficient Distributed Business Management", In proceedings of IEEE International Multi-Topic Conference 2008 (INMIC-2008)

[7] Jeffrey A. Hoffer, Mary B. Prescott, Fred R. McFadden, Modern database management, Sixth Edition, Pearson Education Publishers, Singapore

[8] Thomas Connolly, Carolyn Begg, "Database Systems: A Practical Approach to Design, Implementation and Management," 4th Edition, Addison-Wesley, 2003

[9] William Inmon, Building the Data Warehouse, 2nd Edition, New York: Wiley publisher. Inc, 1996

[10] Rizwana Irfan, Azra Shamim, Madiha Kazmi, Framework for Case Based Object Oriented Expert Warehouse to Enhance Knowledge Management Process for Executive Decision Making, Journal of Computing, Volume 3, Issue 4, May 2011.

[11] Atika Qazi, Azra Shamim, Rubina Adnan, Farooq Azam, A Distributed Data Warehouse Architecture with Fair Query Execution Scheme, ICMLC, 2011

[12] Saif Ur Rehman Malik, Azra Shamim, Zanib Bibi, Sajid Ullah Khan, Shabir Ahmad Gorsi, A Framework for ETL Workflow Management for Efficient Business Decision-Making, ICSCT 2010

[13] Daniel L. Moody, Mark A.R. Kortink, "From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design", In Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) June 5-6, 2000, Stockholm, Sweden

[14] Mohammad Rifaie, Erwin J. Blas, Abdel Rahman M. Muhsen, Terrance T. H. Mok, Keivan Kianmehr, Reda Alhajj, Mick J. Ridley, "Data warehouse Architecture for GIS Applications", In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services (iiWAS '08) , November 2008, Linz, Austria

[15] Data Warehousing and OLAP, www.cs.uh.edu/~ceick/6340/dw-olap.ppt

[16] R. Mattison, "Web warehousing and knowledge management," 1st Edition, New York: McGraw-Hill School Education Group, 1999.

[17] Xin Tan, David C. Yen and Xiang Fang, "Web Warehousing: Web technology meets data warehousing," Science Direct Technology in Society, 2003.

[18] Kai Cheng, Yahiko Kambayashi, Seok Tae Lee and Mukesh Mohania, "Functions of a Web Warehouse," International Conference on IEEE Digital Libraries: Research and Practice, Kyoto, 2000.

[19] Lean Yu, Wei Huang, Shouyang Wang, Kin Keung Lai, "Web warehouse – a new web information fusion tool for web mining," Elsevier, information fusion, science direct, 2006.

[20] Web Data Warehousing, "DVS, web data warehousing", Available: http://www.dvs.tu-darmstadt.de/research/webdataware/, Access Date: 5 December 2008

[21] Yan Zhang and Xiangdong Qin, "Effectively Maintaining Single View Consistency in Web Warehouses," CIT The Fifth International Conference on Computer and Information Technology, IEEE computer Society, pp 199-205, 2005.

[22] Panos Vassiliadis, "Data Warehouse Modeling and Quality Issues," National Technical University of Athens Zographou, Athens, GREECE, 2000.

[23] A Framework for Assessing Database Quality, http://osm7.cs.byu.edu/ER97/workshop4/jh.html, Accessed Date: May 16, 2010.

[24] Informatik V , Matthias Jarke , Matthias Jarke , Lehrstuhl Fur Informatik V , Yannis Vassiliou , Yannis Vassiliou Asdasda, "Data Warehouse Quality: A Review of the DWQ Project", In Proceedings of the 2nd Conference on Information Quality, Massachusetts Institute of Technology, Cambridge, 1997.

[25] Xiaolan Zhu, Susan Gauch, "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web", In Proceedings of the 23rd annual international conference on Research and development in information retrieval, ACM SIGIR, 2000.

[26] Shirlee-ann Knight, Janice Burn, "Developing a Framework for Assessing Information Quality on the World Wide Web", The World Wide Web. Informing Science Journal, 2005.

[27] M. Pamela Neely "Data Quality Tools for Data Warehousing – A Small Sample Survey," Center for Technology in Government University at Albany / SUNY, 1998.

[28] Larry P. English, "Information Stewardship: Accountability for Information Quality," Information Impact International, Inc, 2006.

[29] Peter Block, "Stewardship: Choosing Service over Self-Interest," San Francisco: Berett-Koehler, 1993.

[30] Erhard Rahm, Hong Hai Do, "Data Cleaning: Problems and Current Approaches," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2000.

AUTHORS PROFILE

Umm-e-Mariya Shah is a student of MS(CS) in COMSATS Institute of Information Technology. In addition she is an IT Consultant in Sustainalbe Development Policy Institute, Islamabad. Pakistan. Also she is a visiting lecture in SKANS School of Accountancy, Rawalpindi, Pakistan.

Azra Shamim is working as a research associate in COMSATS Institute of Information Technology, Islamabad, Pakistan. She received her MS(CS) degree from COMSATS Institute of Information Technology, Islamabad, Pakistan.

Madiha Kazmi is working as a Lecture in COMSATS Institute of Information Technology, Islamabad, Pakistan. She received her MS(CS) degree from from National University of Science and Technology, Islamabad, Pakistan.