ISSN 2348 – 9928
IJAICT Volume 4, Issue 5, May 2017 Doi:01.0401/ijaict.2014.07.30 Published on 05 (5) 2017

INTEGRATION OF VERY LARGE HETEROGENEOUS DATABASE FOR MEDICAL DATABASES BY USING DATAWARE HOUSING TECHNOLOGY

Mr. R. Saravana Kumar
Computer Science and Engineering,
Jayam College of Engineering and Technology,
Dharmapuri. Tamilnadu, India

Dr. G. Tholkaippia Arasu
Information and Communication Engineering,
AVS Engineering College,
Salem, Tamilnadu, India

Abstract*— *In medical science and Health care, we need effective tools to acquiring information and systematically analysis large amount of data stored in heterogeneous data base. We have developed and deployed a data warehouse named medical data warehouse (MDW) as a tool to apply integrative approaches to the analysis of large medical data. Integration f a data warehouse enables the proper medical data to provide the required information in a direct, rapid and meaningful way. Medical researchers can view data from various perspectives with reduced query time, thus producing results faster and more comprehensive. Especially, we have organized the framework of data warehouse treated medical informatics combination as different subjects. To get new aspects with especially medical and treatment history, researchers can pick up subjects to analysis with data mining in MDW. We expect MDW will be a useful tool for supporting medical data analysis. MDW should be one of important data sources for Health care management system.

Keywords*— *Medical Sciences, Medical Informatics, Medical Data, Health Care Management.

I. INTRODUCTION

Medical informatics is a scientific discipline which intersects medical related information and computer science. It is a controlled scientific field of study which focuses on acquiring information, storage, retrieval, and processing of medical data to interpret knowledge for the purpose of predication and decision making. It is often difficult to physically model the relationships between constituents, and processing, and final properties. Finally, improved algorithms, higher computer power, and software middleware lead to more effective and easy to adopt data mining technologies.

However the medical data are highly diverse and widely scattered across hundreds of databases in different formats and thus are difficult to query and analyze. So there should be a method to be developed that allows the integration of these medical data stored in heterogeneous databases in a consistent way. Integration medical data between heterogeneous databases is very important for two reasons. The one is, information about a given medical data is often scattered across many different databases: the Patients history records may be stored in one database; the Medicine prescription may be stored in a second database, and so on. But investigating the relations between patients and medicine processing always need medical data the more the best. The other reason is, different databases often contain redundant or overlapping information. Integration data allows cross validation and verification of the databases to identify such information. There are several major integrated technologies including XML, virtual database, database mediation and federation and data warehouse. For providing a method which allows us to gain new insights and a deeper understanding of complex materials systems with integrating, managing, and analyzing these materials data, we have chosen data warehouse technologies as the tool to solve the problem.

II. WHY CHOSEN DATA WAREHOUSE

Many Clinical databases have been constructed by different organizations around the world to support their health care development for more than a decade. Each of these databases may be associated with one or several specific purposes, so the data stored in them are always limited and incomplete. Even the segregation causes many disadvantages including [5]:
It prevents data sharing between materials research;

- It cannot afford enough medical data for data mining;
- Multiple entries for the same data are generated at various locations;
- It causes misunderstandings if the same data at different locations are not updated simultaneously;
- It slows down result-producing process if data obtained from different sources conflict with each other.

And in medical science, researchers should solve several problems before their experiments lead to satisfactory results. How to organize medical data for finding new treatment with desired clinical data? How to query and exploit medical data across the heterogeneous databases? Except for correlations

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.07.30 Published on 05 (5) 2017

between structure data and properties data, how to deal with correlations among processing data, structure data and properties data

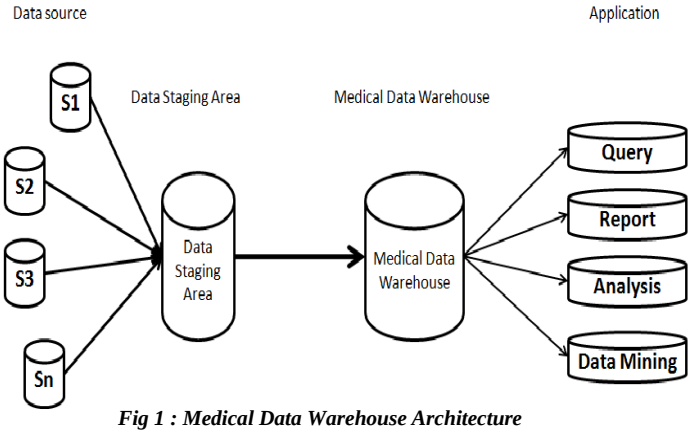

Fig 1 : Medical Data Warehouse Architecture

Because integration of widely distributed data is prerequisite to their analysis, several approaches for data integration have been investigated: linked, indexed data connect flat file databases using the World Wide Web (WWW), or federated database systems which integrate heterogeneous database systems by a central query interface. In contrast, data warehouse systems provide a tight data integration by a common data schema and periodically load all data into a central repository.

The characteristics of data warehouses are very suit for deal with these problems mentioned above. Since the 1990s, data warehouses have been an essential information technology (IT) strategy component. Data warehouses provide the basis for management reports, decision support, and sophisticated on-line analytical processing (OLAP) and data mining. William Innom, who coined the term "data warehouse", defined a data warehouse as "a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management decision" [6]. Furthermore, data warehouse techniques help to overcome two major limitations of local databases: inconsistency of data and time consuming or incomplete queries caused by server restrictions.

Materials data are subject-oriented. Collecting and maintaining data in relations to subjects sound practical and logical. We develop a data warehouse named medical data Warehouse (MDW) for medical science researchers. Our goals were to build MDW:

- To integrate a heterogeneous set of databases related to the clinical data base.
- To created detail tables to store medical and related data, including Medicine perception, Patients details, treatment history and doctor details, etc.
- To provide proper and organized data for researchers to create ad hoc queries, analyses, and reports, particularly data mining, through the development of user interfaces.

III. THE MEDICAL DATA WAREHOUSE (MDW)

The medical data warehouse (MDW) was designed to store Clinical data for long periods of time to support daily inquiry and long term research, and can be accessed from both internet and intranet. For the reason of easily accessing and retrieving data by users, the MDW has been constructed in combining C/S with B/S environment. Its architecture is shown in Figure.1 with four major components: data sources, data staging area, data storage servers, and data access. The data source component includes source databases that supply data to the warehouse. The data staging area is an intermediate database which is used for transferring data from its sources to the warehouse. The data storage servers store the data in the warehouse. The data access component provides an interface for end users to retrieve data, to process, organize or analyze data, and to export data to external environments. MS SQL Server 2005 was chosen for MDW because it supports C/S and B/S, data warehouse applications and also it develops Analysis Services pack. Implementing the MDW involves the two major tasks: (1) creating the warehouse structure with the tables; (2) designing the strategy and tools for populating the warehouse.

3.1. Data sources

The elements found in nature have different informatics associated with those elements such as Electronic Health record, Medical record thermodynamic properties, mechanical properties, crystal structure properties, and others. Compounds have about 39 different properties, such as enthalpy of formation, entropy, etc. The materials properties can be classified into the following categories: element properties: electron affinity, electro-negativity, ionization potential, and atomization potential; and compound properties, which can be broken down into the following: electrical properties, general properties, magnetic properties, mechanical properties, optical properties, structure properties, thermodynamic properties, etc[7] . As mentioned above, the data corresponding to these properties are scattered in different databases on specific purpose.

So to design the medical data warehouse, the first difficulty is to determine the data structure, or what data will be included to the warehouse. Two distinct approaches may be used to determine the corresponding strategy: need-based and availability-based approaches. The need-based approach

examines what data will be needed in the future based on the nature so that these needed data will be collected and be stored to the warehouse. The availability-based approach examines what data: is currently available in the operational systems; and the available data will be selected to the warehouse. Some data to be loaded in the warehouse may not have any immediate use but may find it useful in the future because it is much cheaper to store data than to collect it later.

There are a lot of materials data collected in crystallographic databases (e.g., ICSD[8]) and other data sources including public media. Materials data is generated along its different purposes from experiments to computing. These data may exist in a large variety of formats or even different database manage systems, Oracle or MS SQL Server. Most of researchers submit their crystal data in CIF (Crystallographic Information File). Fortunately, these data are easily to be rearranged and organized in MDW.

3.2. Data warehouse schema

For designing MDW schema, we studied the schemas of each database to be integrated at first. We determined that MDW schema should be as simple as possible. Two major types of data models are widely used for constructing MDW. An Entity-Relationship (ER) model removes the maximum possible redundancies in the data. It provides advantages in transactional processing by making transactions simple and deterministic. On the other hand, dimensional modeling enables speedy access and queries. Although it may use more space to store data, it provides one of the most practical techniques for delivering data to users in a data warehouse.

MDW adopts the star schema data model including one fact table and more than 30 dimension tables. The fact table details general data including medicine, tablet count, chemical combination, etc. The dimension tables describe: the time dimension, the elements dimension which defines its characteristics of the medicine, the dimension which shows the chemical belonged; the space group dimension describes the parameters of the space group, and Medical records that describes the patient information, etc. The details of the dimension tables are not provided in the paper due to a limit on space.

3.3. Architecture

The MDW was developed on Windows 8 and is built as a combining C/S with B/S architecture. The query interface was developed in C#+ASP.NET. For the web-based query interface the IIS (Internet Information Server) is used. MDW is constructed in Microsoft SQL Server 2008. SQL Server 2005 offers database, Online Analytical Processing (OLAP) database, and a set of processes that a system administrator uses to import and maintain data. The data warehouse stores and manages data in the database for performing data mining and analysis reporting.

3.4 Extraction, transformation, and loading (ETL)

Extraction, transformation, and loading (ETL) are key phases of a data warehouse implementation. Populating a warehouse starts with extracting data from its sources. Then, the extracted data must be processed and checked for correctness before it is loaded to the warehouse.

Because materials data bases are often large and may have a relatively poorly defined syntax, load failures are frequently observed and without precautions could result in crashing the load process. For this reason, database loading should be able to recover from errors encountered.

In order to automate a data warehouse population process; an ETL procedure must be developed. Developing an ETL tool may consume about half of the time of a constructing MDW. The ETL tool of MDW has programmed in Visual Studio 2005 IDE in C# with ASP.Net 2.0. The ETL tool are designed to keep loading even in the presence of an error. If partial data has been inserted into the warehouse, an error flag maintained on the related objects is updated to indicate that an error occurred while parsing the object, and that the warehouse entry for the object may therefore be incomplete or contain errors.

IV. CONCLUSION

Our goal is to stores all medical and health care data in a multidimensional data model. Although the medical data stored in MDW is still incomplete. Integrating and querying is only one of the goals to construct the MDW. The most important goal is to do data mining applications based on it. As mentioned above, data mining technologies include some very useful algorithms, such as regressions, neural networks, genetic algorithms, classification algorithms, principal component analysis. These approaches have gained great achievements in different fields including medical and health care and clinical data.

References

[1] C. Ortiz, O. Eriksson,, M. Klintenberg, "Data mining and accelerated electronic structure theory as a tool in the search for new functional materials", Computational Materials Science, 44 (2009) ,1042–1049.

[2] D. Andrew Carr, Mohammed Lach-hab, Shujiang Yang, Iosif I.Vaisman, Estela Blaisten-Barojas, "Machine learning approach for structure-based zeolite classification", Microporous and Mesoporous Materials,117,(2009) 339–349.

[3] J. R. Rodgers and D. Cebon, "Materials informatics," MRS BULLETIN, vol. 31, p. 975-980, 2006.

[4] G. C. Dane Morgan , Handbook of Materials Modeling. Netherland, Springer, 2005.

[5] K.W. Chau, Y. Cao, M. Anson, J. Zhang, Application of data warehouse and decision support system in construction management, Automation in Construction (12) (2002) 213–224.

[6] W.H. Inmon, Building the Data Warehouse(4th Edition), John Wiley and Sons, Inc.,New York, 2007.

[7] R. Hrubiak, L. George, S. K. Saxena, and K. Rajan, "A Materials Database for Exploring Material Properties," JOM, vol. 61, p. 59-62, 2009.

[8] ICSD Inorganic Crystal Structure Database, FIZ Karlsruhe. <http://www.fizkarlsruhe.de/icsd.html>.

[9] R. Ben Mosbah, S. Dourlens, A. Ramdane-Cherif, N. Levy, F. Losavio, "Information management of mechatronic systems materials", International Conference on Computer Systems and Technologies, jan 2011.

[10] Sara Mostafavi and Quaid Morris “Fast integration of heterogeneous data sources for predicting gene function with limited annotation”, Vol. 26 no. 14 2010, pages 1759–1765, May 2010

[11] Kristian Ovaska, Marko Laakso, Saija Haapa-Paananen, Riku Louhimo, Ping Chen, Viljami Aittomäki, “Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme”, volume-2, page-65, 2010.

[12] Ranjit Singh and Dr. Kawaljeet Singh, "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010.