Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KEY BASED APPROACH FOR INTEGRATION OF HETEROGENEOUS DATA SOURCES

[1]KAMSURIAH AHMAD, [1]TENGKU SITI FATIMAH TENGKU WOOK, [2]REDUAN SAMAD

[1]Strategic Management Research Group
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Malaysia
[2] Faculty of Information Technology
Asia e-University, Malaysia

ABSTRACT

Patients' medical records are very important for practitioners to make medical related report or provide knowledge for the management to make a decision. With the advancement of Information Technology, most clinics and hospitals have their own system to process and to store their own medical data. Different clinics or hospitals might use different types of data standards, models, or procedures to process their patients' medical data. The implication of these is the medical data cannot be shared by the systems. A proper mechanism is needed so that the medical data can be integrated and shared among the systems in the hospital. Existing mechanisms on the data integration are not based on keys. In this paper, an integrator tool is proposed to integrate the heterogeneous data sources and mapped into the predefined global schema. The methodology used in this study consists of four phases: Analysis Phase, Design Phase, Implementation Phase, and Testing Phase. The proposed tool is efficient in terms of further improves the results of the integration. Being able to integrate data from different sources and map into a global view will improve the accuracy of information in the database.

Keywords- *Integration, Heterogeneous Data Sources, Global Schema,Medical Records*

1. INTRODUCTION

Systems in the organization are normally domain oriented and exist as separate islands of information. The existing systems are not able to integrate and sharing of information is almost impossible [3]. Most organization face this problem because the systems are build separately and the system does not meant to be integrated at the first place. But when the business grows, data from multiple distributed sources need to be integrated in order to support business decision-making and enterprise management [7]. In real context the application can be as follows: two different patient schemas from different hospitals are going to be integrated in order for sharing of information. To integrate, these two schemas need to be compared and matched, in terms of its structure and semantics to find the similarities and differences. Later a global schema needs to be constructed so that it can represent a common integrated schemas. Given rapidly increasing number of data sources to integrate and due to database heterogeneities, manually identifying schema matches is a tedious, time consuming, error-prone, therefore expensive process [1]. As systems able to handle more complex databases and application, their schemas getting larger, further increasing the number of matches to be performed. Thus, automating this process, which attempts to achieve faster and less labor-intensive, has been one of the main tasks in data integration [10]. However the challenges are, the source schema contain heterogenous structures and semantics, compatible and incompatible among elements. Although many techniques have been revisited or newly developed [5] in the context of schema matching for data integration, but the techniques are not general enough to be applied to other domain.

2. ISSUES IN DATA INTEGRATION

Data integration has recently received considerable attention from researchers in the fields of Artificial Intelligence and Database Systems [2]. The goal of data integration system is to provide an interface for a combination of multitude data sources. Issues with combining heterogeneous data sources under a single query interface have existed for some time. The rapid adoption of databases after the 1960s naturally led to the need to share or to merge existing repositories. This merging can take place at several levels in the database architecture. One popular solution is implemented based on data warehousing [9]. The warehouse system extracts,

Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

transforms, and loads (ETL) data from heterogeneous sources into a single common queriable schema in order to improve the compatibility of data within the systems.

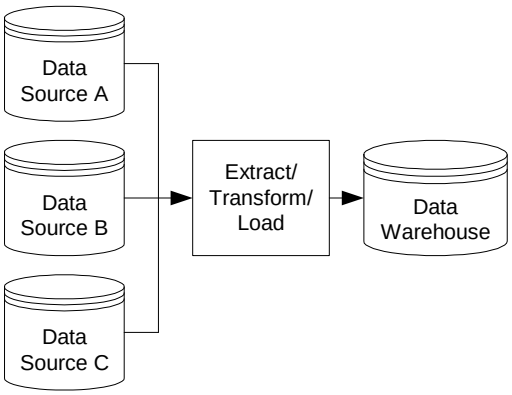

Figure 1: Data Integration Process In Data Warehouse

Figure 1 shows a simple schematic for data warehouse where the ETL process extracts information from the source databases, transforms it and then loads into the data warehouse. The data integration problem is complicated by the facts that
(1) source contain closely related and overlapping data,
(2) data is stored in multiple data models and schemas, and
(3) data sources have delivering query processing capabilities.

Key element in a data integration system is the ability on the language used to describe the contents and capabilities of the data sources. While such a language needs to be as expressive as possible, it should also enable to efficiently address the main inference problem that arises in this context: to translate a user query that is formulated over a mediated schema into a query on the local schemas. As of 2010 some of the work in data integration research concerns the semantic integration problem [4]. This problem addresses not the structuring of the architecture of the integration, but how to resolve semantic conflicts between heterogeneous data sources. For example if two companies merge their databases, certain concepts and definitions in their respective schemas have different or similar meanings. In particular, a data integration system requires a flexible mechanism for describing contents of sources that may have overlapping contents, whose contents are described by complex constraints, and sources that may be incomplete or only partially complete. This paper describes the main languages considered for describing data sources in data integration systems, which are based on extensions of database query languages.

3. MATERIALS AND METHODS

The purpose of this study is to propose a new method in integrating heterogeneous data sources, so that the data is accessible and shareable. Hospital Information System is used as a domain study. In the hospital sector as an example in other domains, many silos applications and disparate data exists and have difficulty in sharing the data. Data contained in the healthcare information system is used by many different units within the hospital. It covers mainly on patient data, management, and other related information. These data are useful for the administrator to make further decision. Healthcare data has no medical sense by itself, but according to a context [2], [8].

Consequently, only data classified as “relevant” is intended to be exchanged and communicated [9]. The data is classified as “relevant” if the data is: i) associated to its production context, ii) deemed useful, and iii) potentially reuseable for the patient, for the hospital, for management purposes, research or social development. However there is a need to overcome the complication of these data during integration. In order to integrate the data from various sources, first the data need to be analyzed to understand its semantic meaning. These are necessary in order to develop the relationship among the integrated data. Due to inconsistency of data, sharing of information is not easy to be done. Based on Ramli [2] the existing hospital information system is not consistent from one hospital to another. Due to this problem, the users including doctors, nurses, or lab assistants need to use different kind of systems to assess data or information. This will lead to slow information searching.

The proposed integration method has three main components: i) the source data, ii) integration process and data mapping, iii) global database schema. Figure 2 shows the proposed process on integration and data mapping in Hospital Information System environment. The idea is to integrate data from different database of similar domain and mapped the integrated data into global database schema so that the query can be done in a single interface. Based on these three components, the methodology used in this study consists of four phases: Analysis Phase, Design Phase, Implementation Phase, and Testing Phase. The steps

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to accomplish this process are as follows: First, data from the source will be analyzed where the data value is compared and evaluated to find similarities and the difference. The data is then mapped to the global schema. The detail steps for each component are discussed.

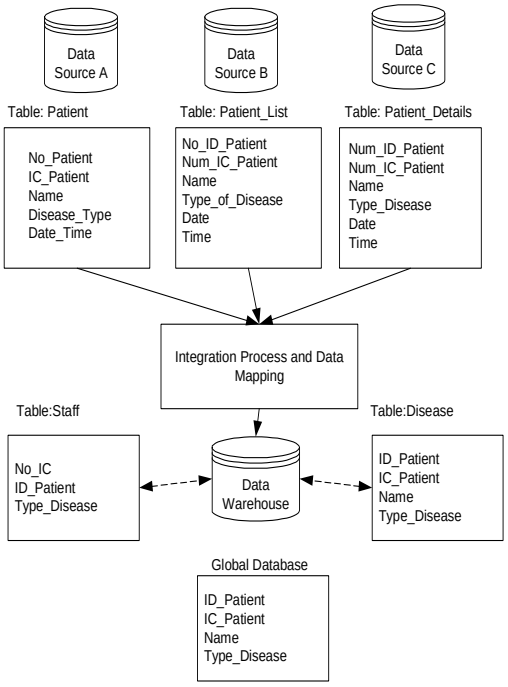

Figure 2. The Propose Integration And Mapping Process

4. ANALYZING THE SOURCE DATA

The data on the same domain but from different data source is reviewed and analyzed. In this study it is assumed that the data is from three different systems. For instance, Systems A has a table named Patient, System B has a table named Patient List, whereas System C has a table named Patient Details. These three table are in similar domain, but have different structures and with different platforms. In order for the data in these three tables to be shared, the data need to be analyzed first to understand the semantic meaning.

5. INTEGRATION PROCESS

The integration technique proposed in this study is based on primary key of each table. Primary keys are a fundamental design feature of relational databases. A primary key is a combination of columns which uniquely specify a row [4]. For example as in Table 1, primary key for all the tables is an attribute called patient ID number where this attribute is unique. However the attribute name may be different from one system to another. For the integration purposes an attribute called patient IC number is used instead, where this attribute is unique for every patient. This attribute is used to find other information about the patient. The patient record on each table is read based on this primary key. There might be the case where the same patient IC number appears in other database. The problem begins when there is a need to integrate the records. If integration is needed then the patient IC number is read and searching process is done to find other number that are similar (if any) from other database. The detail process is described in Figure 2. The next process is mapping the target schema to the global schema. The global schema contains standard predefined schema, where the schema at the source will map to this schema. The global schema is meant to standardize the dissimilarities among the heterogeneous data source schema.

6. RESULTAND DISCUSSION

To evaluate the effectiveness of the proposed method, let us consider the Patient table and its data as displayed in Table 1. From the table it shows that the patient "Mary Edwards" visits three different hospitals, and her record stored in three different databases. From the records, Mary Edwards visits the hospitals because of her three different diseases. If a user queries the data on Mary, then the information about Mary need to be integrated into one database. Therefore a good mechanism is needed so that the information about disease is not duplicated. To approach this, first the IC number of Mary Edward is access to find other records of Mary in other database. If it exists then all the information about Mary will be accessed and integrated. The data are then mapped to the schema at the global.

Table 1 Patient Information

System 1	No_Patient	IC_Patient	Name	Disease Type
	A0102	800808145002	Mary Edwards	Arthritis
System 2	No_ID_Patient	Num_IC_Patient	Name	Type of Disease
	110110	800808145002	Mary Edwards	Dengue
System 3	Num_ID_Patient	Num_IC_Patient	Name	Type Disease
	00296F	800808145002	Mary Edwards	High Fever

The above steps are simplified as follows:

i) The user needs to enter the identification number of a patient (Patient IC number). Based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on this number, the system will search patient record in these three databases.

ii) If the Patient IC number is matched with any patient record in the database, then the patient record will be read and integrated.

iii) These record will then mapped to the global schema.

Figure 3 describes the process of data mapping into predefined global schema. Global schema shows the combination process of data. The result of this process will be displayed on the web browser's screen. The web browser will display the new patient identification number, name and list of diseases.

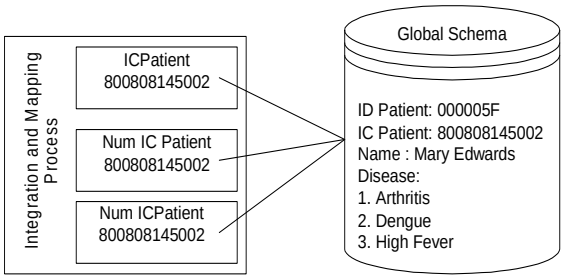

Figure 3 – Mapping To Global Schema

The purpose of this study is to integrate existing systems using semantic as shown in key attribute. This study focused on the hospital system to standardize the patient identification number of each patient from a source system into a targeted system. Patient identity card number is used as a primary key in this development process. Through the process of key based approach, patient identification number is read from a source system and matched with patient identification number in the targeted system. If the number existed in the other system, then a new identification number for the patient will be generated. To test the validity of the algorithm, a running example is shown in Figure 4.

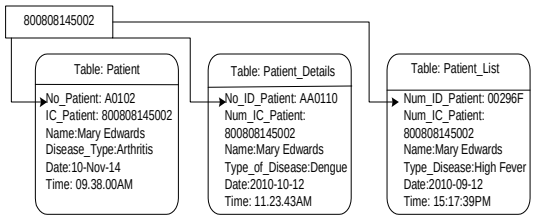

Figure 4: Key Based Integration Method

In this example Patient Identification Number is used as a key with value "80080145002". This value is matched with other value of Patient Identification Number that may exist in other system. If it is matched then all the information about the patient are accessed and a new ID Patient is generated. This information is stored in the global database. Compared with the result from other approaches done by Ahmed et al. [1], Ramli et al. [2], Ding et al. [10], the result shown in this study shows some improvement in terms of accuracy on the data.

7. CONCLUSION

Database integration aims at providing a uniform and consistent view called global schema, over a set of autonomous and heterogeneous data sources, so that data residing in different sources can be accessed as if it was in a single schema. The integration of data sources can be performed in two steps, a matching and a data transformation step. Schema matching, the focus of this paper, is a fundamental operation in the manipulation of schema in formatting match, which takes two schemas that correspond semantically to each other. Manually specifying schema matches is a tedious, time consuming, error-prone, and therefore expensive process, which is a growing problem given the rapidly increasing number of data sources to integrate. This paper presents a technique for matching and mapping health data which consists of three main steps as discussed in the previous section. The formation of global schema is to standardize and to improve the hospital healthcare information management systems. The proposed algorithm is able to integrate all the data that based on primary key without data redundancy. As an immediate task there is a need to find other mechanism for the integration and mapping of data.

REFERENCES

[1] E. Ahmed, Nik Bessis, Yong Yue and Muhammad Sarfraz, "Data Mapping, Matching and Loading using Grid Services", 24th IEEE International Conference on Advanced Information Networking and Applications: 1158-1164. 2010.

[2] F. Ramli, Rosita Mohamed Othman and Nadia Natra Musa, "Developing Conceptual View of Schema Mapping on Date Format for Healthcare Data Warehousing", Proceedings of World Academy of Science, Engineering and Technology 38. 2009.

[3] M. Atay, Artem Chebotko, Dapeng Liu, Shiyong Lu and Farshad Fotouhi, "Efficient schema-based XML-to-Relational data mapping", Information Systems 32(3): 458-476. 2007.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[4] M. Nader. and J. Al-Jaroodi. A Middleware Service for Increasing Applications Integration Availability. Journal of Applied Science 4 (2), 2008, pp. 95-102.

[5] P. Cappellari, Denilson Barbosa and Paolo Atzeni, "A Framework for Automatic Schema Mapping Verification Through Reasoning", Manager: 245-250. 2010.

[6] R. Fagin, Phokion G. Kolaitis Lucian Popa and Wang-Chiew Tan, "Schema Mapping Evolution through Composition and Inversion". 2011.

[7] S. Al-Fedaghi, and E. Al-Dwaisan. Framework for Managing the Very Large Scale Integration Design Process. American Journal of Applied Sciences 9 (2): 2012. 213-222.

[8] T. P. Kuang, Hamidah Ibrahim, Nur I. Udzir and Fatimah Sidi. Security Extensible Access Control Markup Language Policy Integration Based on Role-base Access Control Model in Healthcare Collaborative Environments. *American Journal of Economics and Business Administration.* Volume 3, Issue 1. 2011, pp. 101-111

[9] X. Wang and Xianyi Qian, "A Study of Building Data Warehouse Based on Making Use of Its System Structure and Data Model", International Conference on E-Business and Information System Security: 1-3. 2009.

[10] Y. Ding, Hua Han and Fengming Liu, "Intelligent integrated data processing model for oceanic warning system", Knowledge-Based Systems 23(1): 61-69. 2010.