I.J. Information Technology and Computer Science, **2012, 1, 25-31**
Published Online February 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2012.01.04

Architecture for Accessing Heterogeneous Databases

Mohd Kamir Yusof, Ahmad Faisal Amri Abidin, Mohd Nordin Abdul Rahman
Faculty of Informatics, Universiti Sultan Zainal Abidin, Terengganu, Malaysia
Email: mohdkamir@unisza.edu.my

***Abstract*— This paper presents the architecture for accessing heterogeneous databases. Two major processes in this architecture which are extracting SQL statement and ontology. The algorithms for extracting SQL statement was created and tested in order to improve time performance during searching and retrieving process. Ontology approach was implemented and combined with these algorithms. In ontology approach, web semantic was implemented in order to retrieve only relevant data from database. A prototype based on this architecture was developed using JAVA technology. JAVA technology was chosen because this technology have Jena library. This library is provide API and support SPARQL. Several experiments have been executed and tested. The result indicates this architecture able to improve web query processing in term of time. The result also indicates this architecture able to retrieve and displayed more relevant data to web users.**

***Index Terms*— Heterogeneous Database, Data warehouse, Ontology, Semantic Web, Web query processing.**

1. Introduction

Web query is important element to be considered in architecture application development especially for web based application. The purpose of web based application is to provide the information to web users. In Malaysia, most of people using web based application such as Malaysia Government Portal (http://www.malaysia.gov.my) to access information about education, carrier, tourism, etc. The web users can surffing internet using laptop, notebook, PDA, etc anywhere and anytime to access the information. These information can help web users to make any decision in research, education, etc. Most of web applications stored data into a single database. A single database stored lot of data. Increasing number of data will reduce the space size of storage database. In this case, performance of web query processing become slow to search and retreive lot of data from a single database. In web based applications, three main component invloved; interaction among web user, web application and web server. User interface is provided to web user to enter information (a query) such as education, toursim, etc. The user interface also will display the information/results to web user. Web application server will get a query from web user. This query will manipulate before searching and retrieving process. Then, the web application server will communicate with database server for searching and retrieving process. After data is found, the information/results will display to web user through web user interface. Figure 1 shows the interaction among web user, web application server and database server.

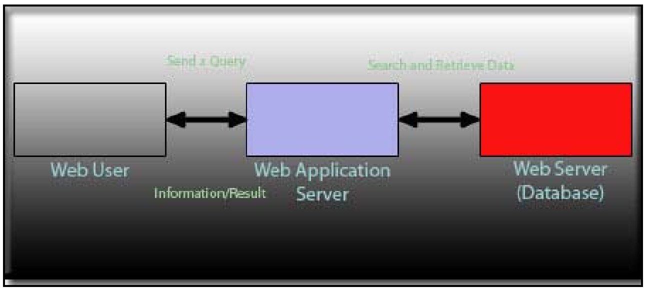


Fig 1. Interaction among Web User, Web Application Server and Database Server

Due to increasing number of data in a single database, a good approach to overcome this issue is heterogeneous databases approach. This approach very suitable implement in web based applicaton especially involved with lot of data. Heterogeneous database means, data is stored in data sources (databases – different type of DBMS) and located at different places. For instances, in heterogeneous databases approach, different data sources are contains different data such as data source1 about tourism, data sources2 about education, etc). In example above, web server will connect to all data sources in order web user can access all data in these data sources. However, issue about web query processing in heterogenous databases still exits caused by difficult to determine what the really users wants [7]. Several research effort has been attempt to solve this problem such as query refinement, etc. These effort is to help the user identify better terms for a query [1][4]. Nevertheless, there is no techniques that enables to identify when the results of the query are relevant [7]. In this research, architecture for accessing heterogeneous databases was designed and implemented in web based applicaton. Experiments in this research focus on two parameters; 1) number of relevance data and 2) response time (in seconds) to access the data.

2. Literature Review

Most of government sector in Malaysia, their web based applications work with heterogeneous database. Hospital is one example using heterogeneous databases approach. Heterogeneous databases system (HDBS) might provide uniform access to electronic patient records in hospital computing environment that uses a MUMPS hierarchical database for storing patient demographic data and a Sybase relational database for storing patient laboratory results [9]. However, the core of the heterogeneous problem in hospital environment is that independently developed and maintained databases are heterogeneous with respect to their query models [9]. Heterogeneous databases approach also has been implemented in geospatial data. Implementation of heterogeneous databases in geospatial data involved three modules; a vector/vector integration module, a raster/vector integration module, and a databases module. Databases architecture was designed to preprocess inputs and to store and export results of the vector/vector and the raster/vector integration steps [8]. The unprecedented increase in the availability of information due to the success of the World Wide Web has generated an urgent need for new and robust methods that simplify the querying and integration of data [2]. A lot of research in the past focuses on developing methodologies for querying heterogeneous data sources. Integration data from existing databases in a distributed environment will give the impact of operations on the databases [14]. One of the approaches in database integration is unified global integration [2]. The purpose of this approach is to facilitate efficient global processing. However, this approach becomes hard to manage as the number and types of data sources increase. Another approach in database integration is system based on mediators and wrapper [2]. This approach is sophisticated that abstract the data sources from the users. The wrapper-mediator is remarkable scalable, and allows the integration of an increasing number of data sources. This research focuses on designing architecture for accessing heterogeneous databases in order to improve web query processing.

3. Data Warehouse

Data warehouse is a data repository that provide variety of data to web users. Data from different sources are located at single place. This single place is called “data repository or data warehouse”. Ideally, web users should be able to access data from the data warehouse without knowing either where data resides or the form in which it is stored [12]. In this research, virtual view approach and datamart approach in data warehouse have been applied. Through this combination approaches, data schema from all data sources will collect and store into data warehouse. These data schema will update automatically once any changing occurs.

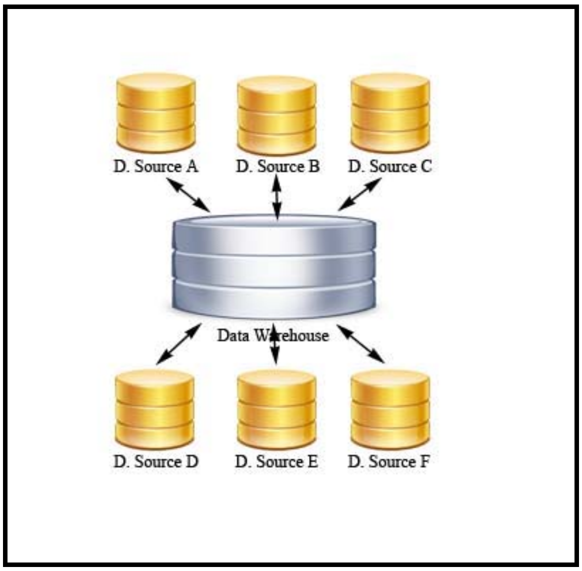


Fig 2: Warehouse Architecture

In figure 2, data schema from six different data sources will extracted and store into data warehouse. Physical data still belonging data sources. Mapping technique will aplply to map data schema to pyhsical data.

4. Semantic Web

The purpose of web based application is to provide information to web users. Most of web based application is designed for humans to read, not for computer programms to manipulate data into meaningful information. Figure 4 shows four major layers involved in the semantics web.

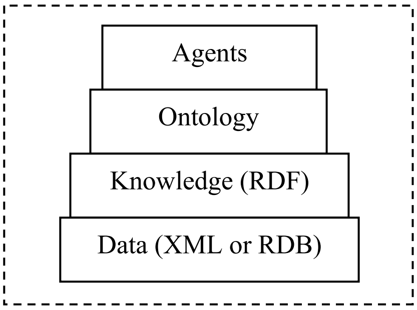


Fig 3: Four major layers in semantic web

XML as data structure language is allows users to add arbitrary structure to their documents. Programmer must know what the document creator means by each structure as XML alone does not capture information about the structures meaning [13]. RDF represent meaning, as users use metadata to describe Web resources and improve in RDF as encoded triples which states that particular resources (subjects) have properties (predicates) with certain value [5]. The second top layers is ontology. Ontology provides a shared and common understanding of a domain that can be communicates between people and heterogeneous and istributed application systems.

5. Ontology in Web Query

In this research, ontology was used in web query processing. The ontology approach was used to retrieve data from heterogeneous databases.

6. Architecture for Heterogeneous Database Access

In this section described about architecture for accessing heterogeneous databases. Two major components in this architecture. First is extracting sql statement from web users and second is ontoloy. Four processes involved in extracting sql statement which are assign the initial query, exploit the initial query, assign any possible query and refinement query. In ontology component, two procces which are semantic mapping and extracting wrapper ontology. This architecture focuses on two number of words for one sentences (such as Data Mining, Data Warehouses, etc). Figure 4 shows the processes flows in architecture for heterogeneous databases access.

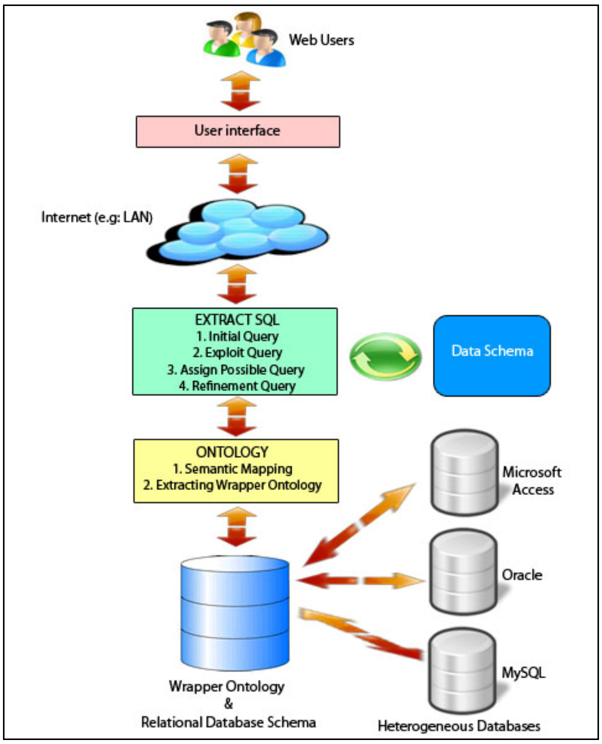


Fig. 4: Architecture for Accessing Heterogeneous Database

6.1 Extraction SQL

Four processes involved which are assign the initial query, exploit the initial query, assign a possible query and possible refinement query.

a) Initial Query

Query from web users through application or system is initial query. For example, web user sends a query about "Political Issues". System automatically assign this query as a initial query.

$$A \rightarrow IQ \qquad (1)$$

Example 1

IQ = Political Issues
A = IQ (Political Issues)

b) Exploit Query

This process will exploit a query from web user. Once web user sends a query about "Political Issues in Malaysia", this section will exploit into four words; Political, Issues, In, and Malaysia.

n = Number of words in A
$M \in \{n_1, n_2, n_3,.., n_i\}$
$M \cup n$

Example 1

A = Political Issues in Malaysia
A_1 = {Political Issues}
A_2 = {in Malaysia}
n = 4
N, number of words in A is 4, where data set in A exploit by space.

Example 2

A = Data Mining
A_1 = {Data Mining}
n = 2
N, number of words in A is 2, where data set in A exploit by space.

c) Assign a Possible Query

In this process, the query will refinement based exploitation query process.

Qn = Number of possible query
$S \in \{Q_1, Q_2, Q_3, .., Q_n\}$

Example 1

A = Data Mining
A_1 = {Data Mining}
Assign any possible query;

α_1 = {Data Mining}
α_2 = {Mining Data}

Number of possible query is 2, where S ∈ {Data Mining, Mining Data}.

d) Refinement Query

Finally, this query will match keywords in data warehouse.

```
Start
Set initial variable is Q
$Q \rightarrow \{Q_1,Q_2,..,Q_n\}$
Loop
            $K = \{K_1, K_2, .., K_n\}$
            If (Qi = = Ki)
            {
              Goto searching($Q_i$,$K_i$);
            }
While Q is null
K ∈ M
Loop
End
```

Fig 5: Matching algorithm

Figure 5 shows the how the algorithm work with get the initial query until match to suitable keywords in data schema.

Example 1

$\beta = \{K_1, K_2, K_3,.., K_n\}$, where K_1, K_2, .., K_n represents as a keywords. These keywords will store in temporary files at server side. Q

Step 1

Match Q_i with any keywords, K_i in temporary files.

Step 2

Hold data sources,DS_i location if found, otherwise keep new keywods, Q_i in temporary files.

Step 3

Repeat step 1 and step 2 until data set Q equal to null.

6.2 Ontologies

Ontology approach was implemented in architecture above. Thr purpose of this implementation is to ensure web users will get a relevant information. Two processes involved in ontology components which are semantic mapping and extracting wrapper ontology.

6.2.1 Build Ontology – Based Query (Extracting Wrapper Ontology)

First step in implementation of ontology is to create a model for query relational database on ontology. Two phases were considered in this model which are offline ontology extraction and online query issuing. In offline ontology extraction, system extracts the explicit classes and relations from the relational schema. In online query, web users can issue semantic query to the system. Figure 6 shows the query relational database on ontology model [10].

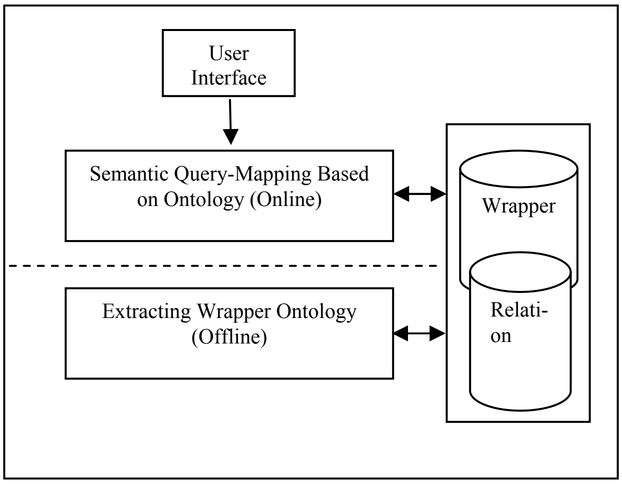

User Interface
Semantic Query-Mapping Based on Ontology (Online)
Extracting Wrapper Ontology (Offline)
Wrapper
Relati-on

Fig. 6: Query Relational Database on Ontology Model

Table 1 below shows the primary key, foreign key and their relationship. The purpose of this table is to show relation among tables.

Table 1: Relational Database for Book Store

Relation	Primary Key	Foreign Key
Book(BookID, Name, Price, ISSN, SupplierID, CategoryID)	BookID	SupplierID,CategoryID
Supplier(SupplierID,BookID, SupplierName, Address, Tel, Email, CSupplierID)	SupplierID	
BookType(BookTypeID, Name)	BookTypeID	
CategorySupplier(CSupplierID, Name)	Csupplier	

Rule 1: If primary key of more than onoe relation is the same is same, merged in one ontological class and their attritube should be merged.

ex: Book rdf:type
rdfs:class
Book(BookID, Name, Price, ISSN, SupplierID, CategoryID)

Rule 2: If the primary of one relation is unique for that relation, and not contain the primary key in another relation, then that relation will be considered as one ontological class.

Ex: BookType rdf:type
rdfs:class
BookType(BookTypeID, Name)
Ex: CategorySupplier rdf:type
rdfs:class

BookType(CSupplierID, Name)

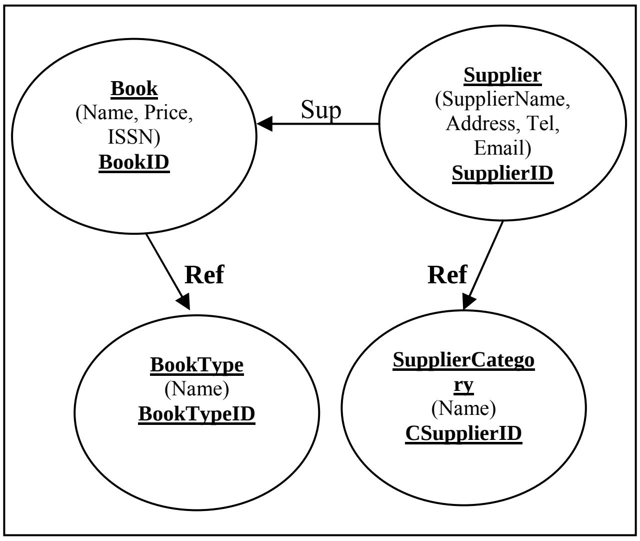


Fig 7: The full wrapper ontology

Table 2 below shows the detailed about ontology.

Table 2: Wrapper Ontology Data

Class	**Property**	**PropertyType**	**Value**	**Related Tables**
Book	BookID, Name, Price,ISSN,Supply	Data Type	Integer String Integer	Book, Supplier
Supplier	SupplierID, SupplierName, Address,Email, Tel	Data Type	Integer String String Integer	Supplier, SupplierCategory
Supplier	Supply	Object Type	Book	
Supplier Category	CSupplierID, Name	Data Type	Integer String	SupplierCategory
BookType	BookTypeID, Name	Data Type	Integer String	BookType

6.2.2 Semantic Query in Relational Database

Semantic query will implement after defining wrapper ontology. The purpose of implementation semantic query is to help user issue semantic query based on extracted ontology concept (based on keywords), and these queries will map onto plain syntatic SQL queries. SPSQL will be used to issue either schema query or data query. Schema query focuses on querying RDF schemas. Data query is related will filter instances. SPARQL-syntax query below shows how to translate into SQL.

SPRSQL-Syntax

```
Select ? name ? price ? ISSN ? Supplier Name ?
WHERE {
?s ex: BookID ? x.
?x ex: name ? bname.
?x ex: price ? bprice.
?x ex: ISSN ? bISSN.
?x ex: Supplier Name ? Sp.
) ? name="Data Mining"
```

This query is to find the book information (name, price, ISSN) where book name ="Data Mining".

SQL Syntax

```
SELECT Book.name, Book.price, Book.ISSN, Supplier
Name FROM Book, Supplier
WHERE Book.SupplierID = Supplier.SupplierID and
Book.Name="Data Mining";.
```

After executing this SQL statement, system will retrieve all related information about data mining from databases and display to web users.

7. Experiment and Analysis

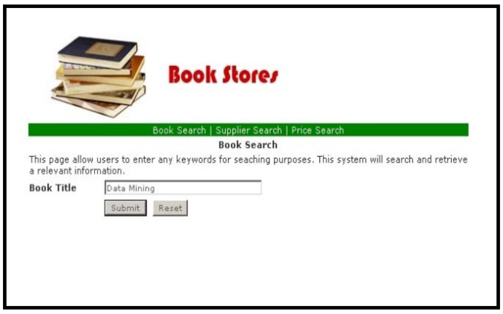
In this section, we describe about implementation, sample query, analysis of proposed architecture, and validation component. The purpose of this section is to produce a prototype based on system architecture was designed in section 6.

7.1 Implementation

The methodology has been implementated in a prototype using JAVA, HTML and four diffferent Database Management System (DBMS). JAVA is programming language to develop a web application system based on architecture in figure 5. JAVA programming was choose because it is a powerful language and make a web application is portable and easily to accessible through World Wide Web. In JAVA, many RDF libararies are provided. The most complete library is Jena. Jena was develop to provide API that was designed specifically for the JAVA programming especially for web application development. Jena also provides ontology API and rules engine for basic inference RDF schemas. It is also support SPARQL by calling ARQ module [10].

7.2 Sample Query

This section illustrates how our system works using a sample query. Suppose that a web user is looking to buy books. Assume that a web user is looking to buy "data mining" book.



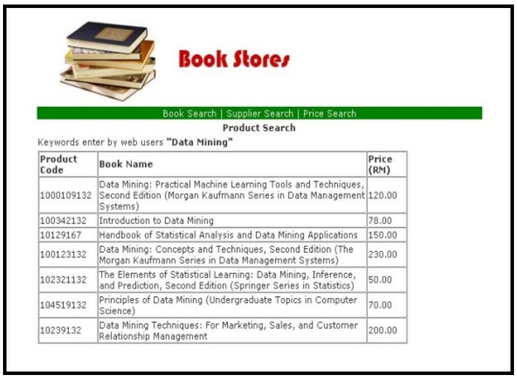
Book Search

This page allow users to enter any keywords for seaching purposes. This system will search and retrieve a relevant information.

Book Title: Data Mining

Fig 8: Search Form

The user interface is allows web users to enter the information (Fig 8). This information is called a keyword. The user requests information through queries submitted to the system via an HTML form. The first process based on architecture in figure 5 is extracting a SQL statement. A SQL statement must through 4 sub processes in extracting SQL statement process. If the keyword is already exists in data schema, this information will submit to next process. The next process is ontology. Two sub processes in the ontology process are semantic mapping and extracting wrapper ontology. In these processes, searching and retrieving process from data sources occurred. The purpose of these processes is to search and retrieve only relevant information to the web users.



Product Search

Keywords enter by web users **"Data Mining"**

Product Code	Book Name	Price (RM)
1000109132	Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)	120.00
100342132	Introduction to Data Mining	78.00
10129167	Handbook of Statistical Analysis and Data Mining Applications	150.00
100123132	Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems)	230.00
102321132	The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)	50.00
104519132	Principles of Data Mining (Undergraduate Topics in Computer Science)	70.00
10239132	Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management	200.00

Fig 9: Result

In figure 9, all relevant information will display to web user. These all relevant information is based on a keyword entered by the web user.

Table 3: Results and Comparison

	Sample Query	**Number of Relevance Data**	**Response Time (Seconds)**
Heterogeneous Database Architecture	Data Mining	20	0.20
	Data Warehouse	40	0.45
	Information Retrieval	15	0.15
	Database Integration	8	0.10
	Education	8	0.13
	University	9	0.14
Proposed Heterogeneous Database Architecture	Data Mining	25	0.18
	Data Warehouse	31	0.33
	Information Retrieval	15	0.12
	Database Integration	8	0.09
	Education	13	0.10
	University	15	0.11

7. 3 Analysis of Proposed Architecture

Table 3 shows the comparison between proposed heterogeneous database architecture and heterogeneous database architecture. This comparison based on number of relevance data and response time (in seconds). Several sample queries have been tested in experiments. Based on table above, the result indicates number of relevance data for "Data Mining" is 20 in both of these architectures. But, response time shows proposed heterogeneous database architecture was decrease about 20% from 0.20 seconds to 0.18 seconds. In "Data Warehouse", number of relevance data for proposed heterogeneous database architecture was deducted from 45 to 34 compared to heterogeneous database architecture. Meanwhile, the response time for proposed heterogeneous database architecture was decreased about 30% from 0.33 seconds to 0.18 seconds. In others queries sample in table 3, such as Information Retrieval, Education and University, the results indicates number of relevance data for proposed heterogeneous database architecture was decreased about 5% to 10% average compared to heterogeneous database architecture. Meanwhile, in term of response time the proposed heterogeneous database architecture was decreased about 10% to 15% compared to heterogeneous database architecture. However, the result indicates number of relevance data for "Database Integration" is same for both of these architectures, but in term of response time the proposed heterogeneous database architecture was decreased about 10% from 0.10 seconds to 0.09 seconds. In conclusion, the result in table 3 indicates implementation of algorithm in extracting SQL statement and ontology approach was improved the web application performance in term of number of relevance data and response time.

8. Conclusion and Future Work

In conclusion, this paper was presented implementation of proposed architecture for accessing heterogeneous databases. This research focuses on improving web query processing. Based on experiments and results above, this architecture is able to improve web query processing in heterogeneous databases access in term of relevance data and response time. The main advantage in this research, the proposed architecture is very efficient and suitable for two numbers of words (such as Data Mining, Information Retrieval, etc). In future work, this proposed architecture will enhance in order it can support more than two numbers of words (such as Introduction to Data Mining, Advanced Research in Computer Science, etc).

References

[1] J.Hartman et. al. Ontology based query refinement for semantic portal, in: integrated publication and information system to virtual information and knowledge environment 2005, 2005, pp. 41-50

[2] Samuel Robert Collins, Shamkant Navathe, Leo Mark. XML schema mapping for heteregeneous database access. Information and Software Technology 44(2002), 251-257.

[3] Eui Kyu Park, Dong Yul Ra, Myung Gil Jang. Techniques for Imrpoving Web Retrieval Effectiveness. Information Processing and Management 41(2005), 1207 – 1233.

[4] B. Velez et al. Fast and effective query refinement, in: Proceeding of the 20th annual international ACM SIGIR Conference on Research Development in Information Retrieval, 1997, pp. 6-15.

[5] Timothy J. Miles-Board. Everything Integrated: A Framework for Associative Writing in the Web", February 2004, University of Southampton.

[6] Jie Cao, Wen Hou, Tingyou Cai. Research of Heterogeneous Database Integration System Based on E-Business, IEEE 2008, 186-189.

[7] Jordi Coness, Veda C.Storey, Vijayan Sugumaran. Improving web-query processing through semantic web knowledge. Data & Knowledge Engineering 66 (2008), 18-34.

[8] Matthias Butenuth, Guido V. Gosseln, Michael Tiedge, Christian Heipke, Udo Lipeck, Monika Sester. Integration of Heterogeneous Geospatial Data in a Federated Database. Journal of Photogrammetry & Remote Sensing 62(2007), 328 – 346.

[9] Walter Sujansky. Heterogeneous Database Integration in Biomedicine. Journal of Biomedical Informatics 34(2001), 285 – 298.

[10] Mostafa E.Saleh. Semantic-Based Query in Relational Database Using Ontology, Canadian Journal on Data, Information and Knowledge Engineering, Vol. 2, No 1, January 2011, pp: 1-16.

[11] Carole Goble, Robert Stevens. State of the Nation in Data Integration for Bioinformatics. Journal of Informatics 41(2008), 687 – 693.

[12] Shi Ming Huang, Tung-Hsiang Chou, Jia-Lang Seng. Data warehouse enhancement: A semantic cube model approach. Information Science 177 (2007), 2238 – 2254.

[13] Jia Lang Seng, I.L Kong. A Schema and Ontology Aided Intelligent Information Integration. Expert System with Application 36(2009), 10538 – 10550.

[14] M. Bright, A. Hurson, S. Pakzad. A taxanomy and current issues in multidatabase system, IEEE Computer 25(3) (1992), 50-60.

Mohd Kamir Yusof obtained her Master of Computer Science from Faculty of Computer Science and Information System, Universiti Teknologi Malaysia in 2008. Currently, he is a Lecturer at Department of Computer Science, Faculty of Infomatics, Universiti Sultan Zainal Abidin (UniSZA), Kuala Terengganu, Terengganu, Malaysia. His main research areas include information retrieval, database integration and web semantics.

Ahmad Faisal Amri Abidin obtained his Master of Computer Science from Faculty of Computer Science and Information Technology, Universiti Putra Malaysia in 2008. Currently, he is a Lecturer at Department of Computer Sciences, Faculty of Informatics, Universiti Sultan Zainal Abidin. His main research areas include computer security, mobile computing and computer networks.

Mohd Nordin Abdul Rahman obtained her PhD in Computer Science from Universiti Malaysia Terengganu in 2008. Currently, he is a Lecturer at Department of Computer Science, Faculty of Infomatics, Universiti Sultan Zainal Abidin, and Terengganu, Malaysia. His main research areas include web services, cloud computing and software engineering.