PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st May 2011. Vol. 27 No.2

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

IMPROVING WEB QUERY PROCESSING THROUGH AN INTELLIGENT ALGORITHM FOR HETEROGENEOUS DATABASES

***MOHD KAMIR YUSOF,FAISAL AMRI ABIDIN, SUFIAN MAT DERIS, SURAYATI USOP**
Faculty of Informatics, Universiti Sultan ZainalAbidin
21300 Kuala Terengganu, Terengganu, Malaysia
E-mail: mohdkamir@unisza.edu.my

ABSTRACT

Performance of web query processing becomes slow caused by increasing number of data. In this paper, intelligent algorithm was created to fix this problem. Four main components involved in this intelligent algorithm are assigning initial query, exploit query, assign to any possible query and query matching. Firstly, web user is needed to enter a keyword then this keyword automatically will assign as an initial query. After that, this initial query must be exploited before assigning process to any possible queries. These possible queries will match to existing query form temporary file that contains data schema and map to only select data sources. In this methodology, XML will use for mapping to heterogeneous databases. XML is really effective for mapping to heterogeneous databases. This methodology has been implemented in a prototype and applied to web queries. Another issue in with web query is difficult to search for information that best reflects the user's need information. According this problem, intelligent algorithm was created and tested by developing simple application based on system architecture. This intelligent algorithm that was good for fixed problems by improving web query processing for heterogeneous databases. The intelligent algorithm was implemented and tested for heterogeneous database environment.

Keywords: *Intelligent algorithm, Web query processing, XML, Heterogeneous database*

1. INTRODUCTION

Website is an important tool for web users to gain information such as education, entertainment, health, etc. The information on the World Wide Web has lead to the need processing intelligently to address more of the user's intended requirements than previously possible [1]. The purpose of web query is to search for information that best reflects the user's need information. However, sometime website cannot produce or display relevant results to web users [1]. This problem occurs caused by less intelligent algorithm in web query engine. In this research, intelligent algorithm was created to improve web query processing in order to display possible relevant information to web user. Intelligent in our context mean, manipulate a query from web users to any possible queries. In query manipulation process, the query must be exploited and combined again to produce any possible query. These possible queries will match with special keywords in a temporary file. The temporary fileallocated lots of different keywords. Through this file also, these keywords maps to data sources. Through this mapping, system automatically refers directly to selected data sources for searching and retrieving process. Keywords in temporary file will update automatically once any updating process occur in certain data sources.

A book inventory system was developed based on architecture that contains intelligent algorithm. The purpose of this development is to test a performance of query processing involved with heterogeneous databases.

2. HETEROGENEOUS DATABASE

This research is ongoing efforts to develop more intelligent and robust methods for querying and integrating data from heterogeneous data

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

sources. The unprecedented increase in the availability of information due to the access of the World Wide Web has generated an urgent need for a new and robust methods that simplify the querying and integration data [2]. Based on the past researches, researchers focus on developing methodologies or technique for databases integration. The purpose is to integrate data from existing databases in a distributed environment while minimizing the impact of operations on the databases [3]. One of the approaches is to use a unified global integration schema, such as the relational schema, to facilitate efficient global processing. This approach is efficient for web query and data integrate but their global schemas become hard to manage as the number and types of data sources increase [2]. Another approach for database integration is mediators and wrapper. This approach is remarkably scalable, and allows the integration of an increasing number of data sources. Mediation does not store any data on its own rather it provided a virtual view of the integrated sources [4].

3. DESIGNING INTELLIGENT ALGORITHM

In this section, structure for an intelligent algorithm was described. This structure was designed in order to integrate with system architecture before implementation phase. Five core processes in this intelligent algorithm are initial query, exploit a query, assign possible queries, assign keywords (mapping to data source), matching.

3.1. Initial query

Initial query is query received by system from web users. Through this intelligent algorithm, assign initial query equal to X.

X → Initial Query, for example web user request about "Data Mining". If request equal to "Data Mining", so X → Data Mining.

Y ∈ $\{X_i, X_{i+1}, X_{i+2} \ldots X_n\}$, where i = 1,2,3, n

Number of initial query (Y) is depending on number of query send by web users.

3.2. Exploit a query

An initial query will exploit one by one. The process of exploitation is based on algorithm below:

```
Start
assign x and i
x = "Data Mining"
i = exploit[x," "]
for (j=0; j<count(i); j++) // looping process until data
equal to null
{
   m = i[j]
   n = n.m
}
End
```

Fig. 1: Exploit initial query algorithm

Example

Based on algorithm above, consider initial query, X →"Data Mining". The results afterexecuted this algorithm is i[0] → "Data" and i[1] → "Mining". However, this result can write as X ∈ {Data, Mining}.

3.3. Assign Possible Queries

After exploit process, these data will combine in order to produce any possible queries. The algorithm to produce any possible queries was created as below:-

```
Start
assign a,b
for (i=0; i<count(n); i++)
{
   a =a." ".a[i]
   for (j=0; j<count(TempFile); j++) // looking to
       temporary file
   {
     b = b →b[j]
     a ≠ b
     i ≠ j
   }
}
End
```

Fig. 2: Assigning any possible queries algorithm

Example

Based on algorithm above consider an initial query, X→ {Data, Mining} have been exploit. The results produce two possible queries; "Data Mining" and "Mining Data".

m[0] → Data Mining
m[1] → Mining Data

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st May 2011. Vol. 27 No.2

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

3.5. Matching, Search and Retrieve Data

This process is important to ensure only relevant data will select from different data sources. The most important in these processes are to match keywords in temporary file match among possible queries. Below is algorithm for matching, searching and retrieving data process from different data sources.

```
Start
assign a,b,c
y ∈ {x_i, x_{i+1}, x_{i+2}, x_{i+n}} // i equal to 0, 1, 2, …, n
for (i=0; i<count(x),i++) { // x represent number of
      possible queries
fileopen (tempFile) // temporary file contains list of
      keywords
  if (y[i]==tempFile[i])
    go to mapping data schema
  else if (y[i]!=tempFile)&&(y[i]==data[i])
    store new keywords and data schema // refer to new
      data sources
  else
    loop until y equal to null
}
End

Store New Keywords ()
{
A → keyword
Assign number[i] of keywords
Set a data schema and data source (physical data)
Loop until data not found in data source
σ → {Data Schema}
β → {DS1, DS2, DS3, DSn}
Therefore,
A →σ →β
Store at temporary file in server
}

Mapping ()
{
 // Mapping to data schema
  Assign a, b
 A is keyword
A → data schema → data source (physical data) → data
D.S ∈ {D1, D2, D3, Dn} // set of data schema
D.SR ∈ {} // set of data source, mean number of selected
      data sources
Search and Retrieve () // go to this function
D → Information // map to only relevant data
}

Search and Retrieve ()
{
 // searching and retrieving process
D is data source
M is data
D → M
D ∩ M
D ∈ {D_i, D_{i+1}, D_{i+2}… D_{i+n}}
}
```

Fig. 3: Matching, Searching and Retrieving Algorithm

Example

Suppose we have 3 possible queries, $M \equiv \{x_1, x_2, x_3\}$. Firstly, find and match these queries in temporary file. Matching process will loop until number of possible queries equal to null, $M \neq$ and $M \geq 1$. If $M \leq 0$, hence number of possible queries is null. In this theorem, keyword (x_i) will store with data schema (y) and data source (destination, z). So, we can write $X_i \rightarrow Y \rightarrow Z$. In this case, if X_1 equal to Xi, hence go directly to specify data source. In second cases, if X2 equal to Xi, hence go directly to specify data source. Otherwise, find and match again until number of keyword (x_i) equal to null. If not found, we assign a new keyword, x_i and store into temporary file. Next process is to assign a data schema and data source.

3.6. Display Result/Information

The last process in this intelligent is to display information to web users based on possible queries were assigned. We can assign, $H \equiv \{d_1, d_2, d_3, d_n\}$, where d1, d_2 until d_n are set of information.

4. PROTOTYPE ARCHITECTURE AND IMPLEMENTATION

The methodology has been implemented in a prototype using J2EE technologies. Simple application was develop using JSP (Java Server Pages) based on architecture in figure 4.

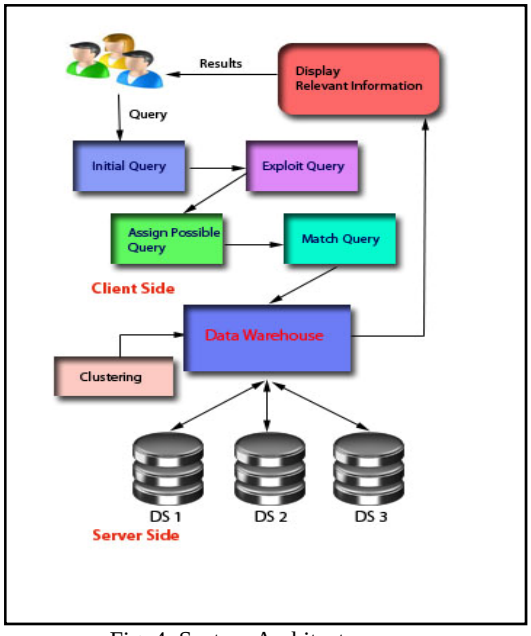

Fig. 4: System Architecture

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st May 2011. Vol. 27 No.2

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

The system architecture in figure 4 is divided into two parts; client side and server side. The client is a web browser that presents the web pages to web users. The purpose of this web page is to gather query from the web users and present the query results (relevant information). Four components contains in client side: (1) initial query, (2) exploit query, (3) assign possible query and (4) match query. The process for each component already explained in section 4. Meanwhile, in server side data warehouse approach was chosen to implement in this architecture. The main function data warehouse is to collect and store keywords and data schema. The purpose of this function is to map to selected data sources. Intelligent algorithm was created to retrieve only relevant data before display the relevant information to web users. This intelligent algorithm was designed to help web users to get only relevant information.

4.1. Implementation

The prototype is implemented as a web application using JSP (Java Server Pages). This environment was chosen because it would make the system portable and easily accessible through the World Wide Web (WWW). Meanwhile, J2EE (Java 2 Platform Enterprise Edition) was chosen as a platform for server programming in Java programming language. The J2EE platform simplifies enterprise applications by basing them on standardized, modular components, by providing a complete set service to those components, and by handling many details of applications behavior automatically, without complex programming [5]. This platform also supports JavaBeans components, Java Servlets API, JavaServer Pages and XML technology. XML is standard for data exchange on the World Wide Web [2][6]. Figure 5 shows an enterprise application model involves with J2EE platform. In this model, 3 components are divided into client-side presentation, server-side business logic and enterprise information system. J2EE is the middle part between client side and enterprise information system.

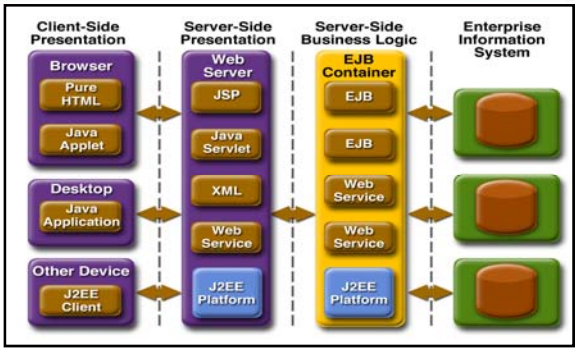

Fig. 5: Enterprise Application Model

4.2. 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.

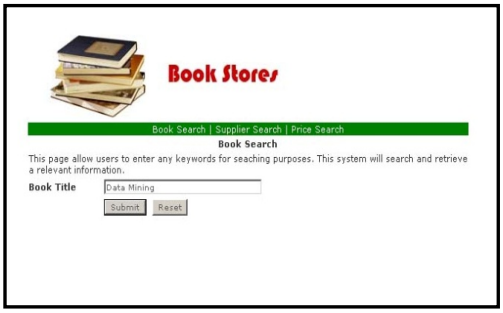

Figure 6: Search Form

The user interface is a Web site which is allows the web users to enter initiate keywords (Fig 6). The user requests information through queries submitted to the system via an HTML form. Once the query has been submitted, it is sent to the client side. Once this query has been submitted, this query will define as an *initial query*. After that, this initial query will exploit into two words; data and mining. Then, the next process is refinement possible query. This process will refine and display any possible query, for instance "data mining" or "mining data". Last module in client side is matching process. This process will communicate to data warehouse in order to find and match a possible query with keywords in data warehouse.

Meanwhile, on server side, keywords in data warehouse have been assigned to data schema. The code processes in client side are searching and retrieving. In this concept, data schema was stored in data warehouse, but physical data store in data sources. Once any changing occurs in data sources, automatically data schema in data warehouse also changed. Based on the data schema, system directly

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st May 2011. Vol. 27 No.2

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

maps to certain data sources. System automatically retrieves any possible and relevant information.

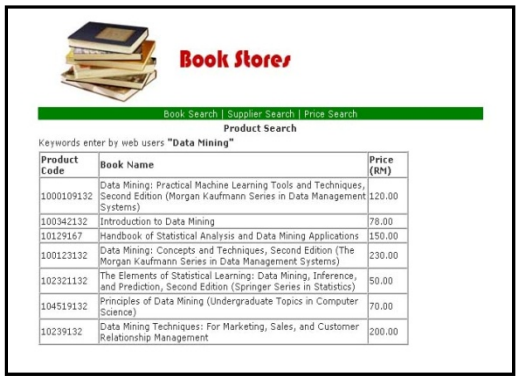

Figure 7: Result

In figure 7, all relevant information will display to web user. These all relevant information is based on keywords entered by web user and through processes or modules on client side and server side.

4.3. Validation

In section 5.2, sample queries ("Data Mining") have been executed with the results shown in Figure 7. Based on figure 7, only relevant information will display to web users.

Table 1: Query Result

Sample Query	Number of Relevance Data	Assign to Possible Queries	Response Time
Data Mining	4	(Data Mining) or (Mining Data)	2 sec
Introduction to Database	5	(Introduction to Database) or (Database)	2 sec
Black Magic	8	(Black Magic) or (Magic Black)	3 sec
Data Structure	12	(Data Structure) or (Structure Data)	3 sec
Advanced Database	3	(Advanced Database) or (Database Advanced)	1 sec
Information Retrieval	5	(Information Retrieval) or (Retrieval Information)	2 sec
Database Integration	2	(Database Integration) or (Integration Database)	1 sec

Table 2: Query Result

Sample Query	Number of Relevance Data	Query	Response Time
Data Mining	15	Data Mining	5 sec
Introduction to Database	8	Introduction to Database	3 sec
Black Magic	22	Black Magic	6 sec
Data Structure	23	Data Structure	7 sec
Advanced Database	24	Advanced Database	6 sec
Information Retrieval	7	Information Retrieval	3 sec
Database Integration	9	Database Integration	3 sec

Table 1 shows the number of sample queries have been executed and the number of relevance data was display to web users. These sample queries show that the addition of intelligent algorithm helps to improve query results. The main function intelligent algorithm in this implementation is to retrieve only relevant information to web users. In order to retrieve only relevant information, the initial query entered by web users must be through 4 components; 1) initial query, 2) exploit query, 3) assign possible query and 4) match query. Table 1 also, display the time for searching and retrieving process. The performance for searching and retrieving process also is good. Meanwhile table 2 shows the numbers of sample queries have been executed without intelligent algorithm. Based on results table 1 and table, the results indicated intelligent algorithm is needed to improve performance for web query processing.

5. CONCLUSION

In this paper a methodology for improving web query processing was designed with designing intelligent algorithm for searching and retrieving data from heterogeneous databases. Four major components in this intelligent algorithm are assigning initial query, exploit query, assign possible query, matching query and illustrated them with examples implementation. This methodology has been implemented in a prototype and applied to web queries. The results indicated the intelligent algorithm was designed be able to improve web query processing.

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

REFERENCES

[1] JordiConesa, Veda C. Storey, VijayanSugumaran. (2008). Improving web-query processing through semantic knowledge. *Data & knowledge engineering*, 66, 18-34.

[2] Samueal Robert Collins, ShamkantNavathe and Leo Mark (2002). XML schema mapping for heterogeneous database access. *Information and software technology*, 44, 251-257.

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

[4] MajidKazemian, BehzadMoshiri, Hamid Nikbakt and Caro Lucas (2005). Architecture for Biological Database Integration. *AIML 05 Conference*, 19-21 Dec 2005, Cairo Egypt.

[5] Askar S. Boranbayev. Defining methodologies for developing J2EE web-based information systems. *Nonlinear analysis* 71(2009), e1633 – e1637.

[6] Java 2 Platform, Enterprise Edition (J2EE) Overview. http://java.sun.com/j2ee/overview.html

AUTHOR PROFILES:

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

Ahmad Faisal AmriAbidinobtained his Master of Science Computer 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.

Nor SurayatiMohamadUsopobtained her Master of Science Computer from Faculty of Computer Science and Information Technology, Universiti Putra Malaysia in 2009. Currently, she is a Lecturer at Department of Information Technology, Faculty of Informatics, Universiti Sultan ZainalAbidin, Terengganu, Malaysia.

MohdSufian Mat Derisobtained her Master of Education (educational technology) from Faculty of Education, UniversitiTeknologi Malaysia in 2006. Currently, he is a Lecturer at Department of Multimedia, Faculty of Infomatics, Universiti Sultan ZainalAbidin, Terengganu, Malaysia.