Journal of
Information Technology & Software Engineering

Yusof et al. J Inform Tech Soft Engg 2011, 1:1
DOI; 10.4172/2165-7866.1000102

Research Article Open Access

Improving Web Query Processing through Mapping Algorithm to Data Warehouse for Bioinformatics

Mohd Kamir Yusof*, Ahmad Faisal Amri Abidin and Mohd Sufian Mat Deris

Faculty of Informatics, Universiti Sultan Zainal Abidin, Terengganu, Malaysia

Abstract

Bioinformatics is combination between biology and informatics. Most of biological data is used by private or government sectors such clinics, hospitals, etc. Web based application is tool to access this biological data. A good methodology or technique is needed to store huge of biological data. In this research, two issues have been identified. First is data integration and second is web query processing. In this paper, technique for data integration and mapping algorithm will propose in order to improve web query processing in bioinformatics. Data integration means integration different data sources and store into a single data source. This single database source will store only important information such a keyword and data source destination. The keyword in data warehouse will match will information enter by web user. Then, a mapping algorithm will map to only relate data source for searching and retrieving process. A simple web based application was developed and tested. Several experiments have been done and results indicates a mapping algorithm and data warehouse approach for data integration able to improve web query processing in bioinformatics in term of time performance.

Keywords: Bioinformatics; Data Integration; Data Warehouse; Mapping Algorithm

Introduction

Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical chemistry) and applying informatics techniques (derived from disciplines such as applied math, computer science and statistic) to understand and organize the information associated with these molecules, on a large scale [1]. The aims of bioinformatics are threefold. First, at its simplest bioinformatics organizes data in a way that allows researchers to access existing information from different data sources (protein data bank for 3D macromolecular structures) [2,3]. The second aim is to develop tools and resources that aid in the analysis of data. The third aim is to use these tools to analyze the data and interpret the results in a biologically meaningful manner. Bioinformatics is based on wealth of diverse, complex and distributed data resources. Challenge in bioinformatics is data integration. This challenge occurs because of growing number of resources [4]. Nowadays, bioinformatics sources are important for medical center, education institution, etc. Normally, different bioinformatics sources are allocate at different places. The issue in bioinformatics is data integration. How to integrate data from different sources? Aims bioinformatics above cannot be achieved if no solution or suitable method will apply to solve issue above.

Overview

This section will explain the overview of data integration and data warehouse approach.

Data integration

Data integration is integrated data among several of data sources from different sources [5]. Integration is necessary due to the large, and increasing, number of data resources within bioinformatics. Based on annual *Nucleic Acids* Research journal database supplement listed 96 databases in 2001 [5] and 800 + in 2007 [6]. A question for bioinformatics database integration is "why so many data resources?" Here are some reasons: There is an ecosystem of primary data collection feeding secondary and tertiary databases. The Web makes it (too) to easy to publish. There are many types of data and each has its own communities and its own repositories. Each new sub-disciplines its own biases. The distributed and diverse nature of the discipline promotes a "long tail" of specialist resource suppliers rather than the few centralized data centers such as particle physics [2] 4. Consequently, each type of data has a multiplicity of resources, many replicating, partially overlapping or presenting slightly different view on more or less the same data types. For example, there are some 231 different pathway resources, but this number seems excessive. It appears that is easier, more desirable, or more expedient, to create a database afresh than it is adapt or re-use existing resources. Data integration for bioinformatics is needed to help scientists in "data surf". A wide variety of technologies, techniques and systems have been explored and exploited over the past 15 years. Data integration system must be developed using suitable method or technique in order to integrated data from different sources. Normally, bioinformatics have lots of data. These data are allocated at different places (database). The purpose of these databases integration is to help users or scientist surfing all bioinformatics data. Once users key in a keyword for searching certain information, system automatically search and retrieve relevant data from different sources. However, to produce a good system for bioinformatics data integration, a suitable approach must be identified before development process.

Data warehouse

Data warehouse is a global repository that stores preprocessed queries on data, which reside in multiple, possibly heterogeneous, operational query base for making decision effective decision. The content of a data warehouse may be replica part of some source of data or they may be results of preprocessed queries or both. This method of data storage provides powerful tool-helping project organizations in making decision [7]. A data warehouse (DW) is very successful

***Corresponding author:** Mohd Kamir Yusof, Faculty of Informatics, Universiti Sultan Zainal Abidin, Terengganu, Malaysia, Tel: 609-6653302; Fax: 609-6673412; E-mail: mohdkamir@unisza.edu.my

Received August 04, 2011; **Accepted** November 17, 2011; **Published** November 19, 2011

Citation: Yusof MK, Abidin AFA, Deris MSM (2011) Improving Web Query Processing through Mapping Algorithm to Data Warehouse for Bioinformatics. J Inform Tech Soft Engg 1:102. doi:10.4172/2165-7866.1000102

to many business organizations. The key factors for this successful is where this approach allow the storage and analysis of large of amount of structured business data [8]. DW is also called "multidimensional" data model, where important to business events, e.g., sales, are modeled as so-called facts, characterized by a number of hierarchical dimensions, e.g., time and products, with associated numerical measures, e.g., sales price. The multidimensional model is unique. This model is providing a framework that is both intuitive and efficient. This model also is allowing data to be view and analyzed at the desired level of detail with excellent performance. Traditional data warehouse have worked very well for traditional, so called structured data, but recently enterprises have become aware that DWs are in fact only solving a small part of their real integration and analysis need. Data Warehouse approach has a good potential to improve with adding a new intelligent algorithm in order to improve system performance. The proposed method will describe in section 3.

Proposed Methodology

A methodology is proposed in this paper. In this methodology, a mapping algorithm will create and data warehouse approach will apply for data integration.

Mapping algorithm

The purpose of this mapping algorithm is to map a suitable a data sources based on query requested by users. Once system receive query from users, system automatically map to certain data sources. In this case, system no need go to one by one data sources to find related data. This new intelligent algorithm can improve query efficiency and response time between users and system. Based on Figure 1, an application will communicate with a single database. A mapping algorithm is need to make a searching and retrieving process become more effective. This is important to ensure web users get relevant information. Based on Figure 2, mapping algorithm was created. The purpose of this algorithm is to map a selected data sources based on a keywords entered by web users. Through this algorithm, searching and retrieving process will only occurs at selected data sources. After that, a result will display to web users.

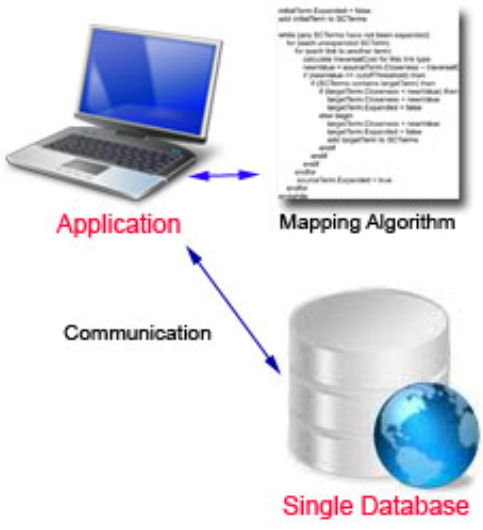

Figure1: Apply Mapping Algorithm.

Asssign Variable
Data = A_i
Key = K_i
Query = Q_i
Data Sources = D_i
Start
$A_i = M_i.D_i$
Loop $Q_i \rightarrow M_i$
(until data is null)
$S \leftarrow Qi$
$M_i \rightarrow D_i$
$R \rightarrow$ Result (Display result to users)
End

Figure2: A Mapping Algorithm.

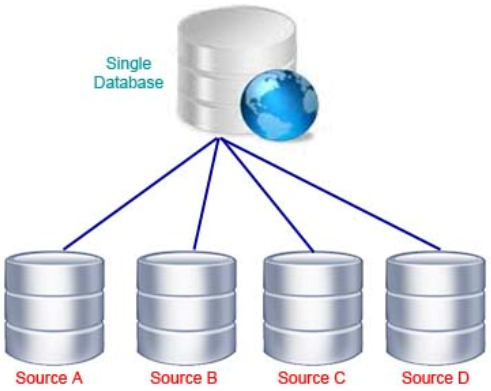

Figure3: Designing Warehouse Approach.

Figure4: A Architecture for Biological Data Integration.

Data Warehouse

A concept for warehouse is collect and allocated all data from different sources into a single place. However, in this paper, warehouse approach is apply, but not all data from different sources is collect, but only certain data with assigned key is collected and put into a single place (a single database). Based on Figure 3, only certain keywords from different data sources will store into a single database.

Integrated between mapping and data warehouse approach

Figure 4 shows a new architecture for biological data integration.

In Figure 4, web users are needed to enter any keywords. After that, web application will communicate with single database. The single database has a relation with registered data sources. Mapping algorithm will map a certain data sources based on keywords enter by web users. After mapping stage, searching and retrieving process will be executed. Finally, a result will display to web users.

Experimental Results

Simple application has been develop to test capability of combining a mapping algorithm and data warehouse approach for accessing biological data.

Experiment 1

In experiment 1, data from different source will integrate using data warehouse approach. Keywords for each data from different data sources will assign and store into a single database. This single database is called "data warehouse". Figure 5 shows interface for bioinformatics application.

Figure 6 shows interface for searching. Web user is needed to enter information to text field, etc. After that, user is needed to press button "submit". After system receive query from user, the system will communicate with "data warehouse". The system will match a keywords entered by users and keywords in data warehouse. After that, system will search and retrieve selected databases based on keywords. Finally, a result will display to web user. Figure 7 shows interface after searching and retrieving process. The results will show to web user.

Experiment 2

In experiment 2, several queries or information was entered by web user. This experiment shows comparison between mapping algorithm and without mapping algorithm in term of time performance for web query processing. Table 1.

Figure5: Interface for Bioinformatics Application.

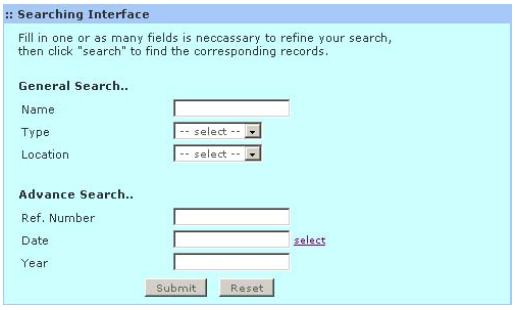

Figure6: Searching Interface.

Figure7: Interface after searching and retrieving process.

Query/Information	Time Performance (Seconds)	
	Mapping Algorithm	*Without Mapping Algorithm*
Pituitary	5s	8s
Optic Chiasma	7s	11s
Pineal Gland	3s	5s
Hypothalamus	4s	6s
Posterior	5s	8s
Anterior	3s	7s
Cerebellum	3s	7s
Spinal Cord	2s	4s
Cerebrum	3s	7s
Adrenal Cortex	4s	8s

Table 1: Comparison between Mapping Algorithm and Without Mapping Algorithm.

Conclusion

As a conclusion, this proposed methodology can be implemented for biological data integration. The data integration for biological data is needed to help users to get relevant information about bioinformatics. This methodology focuses on bioinformatics domain. However, this methodology can be implementing in others domains for future work.

References

1. Luscombe NM, Greenbaum D, Gerstein M (2001) What is Bioinformatics? An Introduction and Overview. Yearbook of Medical Informatics 83-92.
2. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, et al. (1977) The Protein Data Bank. A Computer-Based Archival File for Macromolecular Structures. Eur J Biochem 80: 319-324.
3. Berman HM, Westbrook J, Feng Z (2000) Gilliland Data Bank. Nucleic Acids Res 28: 235-242.
4. Goble C, Stevens R (2008) State of The Nation in Data Integration for Bioinformatics. J Biomed Inform 41: 687-693.
5. Baxevanis AD (2001) The Molecular Biology Database Collection: an updated compilation of biological database resources. Nucleic Acids Res 29: 1-10.
6. Galperin MY (2007) The Molecular Biology Database Collection: 2007 Update. Nucleic Acids Res 35: 3-4.
7. Chau KW, Ying Cao, Anson M, Jianping Zhang (2002) Application of Data Warehouse and Decision Support System in Construction Management. Automation in Consortium 12: 213-224.
8. Torben Bach Pedersen (2009) Warehousing The World: A Vision for Data Warehouse Research. LNCS. Springer 3: 1-17.