# COMPARATIVE STUDY BETWEEN DATA WAREHOUSE AND MEDIATION APPROACHES IN INFORMATION INTEGRATION OF HETEROGENEOUS BIOLOGICAL DATA

SUHAILI BINTI BEERAN KUTTY

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

OCTOBER 2009

# CHAPTER 1

# PROJECT OVERVIEW

This chapter covers about problem background of the research and come out with statements of the problems, objectives, scope, and importance of the research study.

## 5.8 Introduction

In recent years, there are various types of data source in different format available in such field that we call heterogeneous data. This project will focus on heterogeneous biological data which is has diversity of sources data in a form of files and databases. Biological data sources store life sciences information that get from several finding techniques like experiments or computational technologies and the data consists of various file format such as flat file, XML and pdb file and also cover in different data types such as genome, proteome metabolic pathways and

regulatory pathways[1]. It is difficult to manage these several heterogeneity sources and these sources work dependently where biologist or researchers need to access them one by one to complete their tasks and experiments. It is time consuming and because of that, information integration approaches have been introduces where several databases can be made to work together. Three principal approaches of information integration are federation, warehousing and mediation [2] where each of this has their own advantages and disadvantages. Focusing on combining of warehousing and mediation approach that a complementary of advantages and disadvantages of both approaches. Response time of retrieval data in mediation, warehousing and combination of both approaches is subject to be analyses.

## 5.9 Background of the Problem

Biological data is a data or measurement collected from biological sources such as organelle or other natural products and commonly stored in database or file systems. Biological databases contain valuable information covering on various research areas including genome, protein, metabolic pathway and microarrays. Biologists use a large number of biological databases to find relevant information for their research [3].

There are exponential growths of existence biological databases such as Swiss-Prot, PIR and PROSITE on protein domain and examples genome databases are Genbank, DDBJ and EMBL Nucleotide DB. Data store in different type of data models such as flat file model, relational model and object-oriented model [3]. These databases also located at different locations and have different query capabilities like full-text searching.

This heterogeneity of data sources need to be integrated to provide researchers with a service where they can easily access and retrieved relevant information. Accurate information will support their hypothesis of the research and also valuable in validation process. Information integration has many approaches but the tree common approaches are federated database, mediation approach and data warehouse approach.

Federated database approach is the simplest approach where all the sources are independent and one source can call on others to supply information. Second approach is using mediator as a medium to integrate the data. This approach does not store any data, but using mediator to integrate various sources depends on user queries before display result to the user. Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) is example of the mediator-based integration system [4]. The last common approach is warehouse integration where the several of data are located into a local warehouse.

Commonly, the warehouse approach is better in optimization of query because the data locally store in global database and users can add their own annotation to the data. The problem arises when this data are not updated and besides that this data need to be converted in some specific format before it can be store to the global database.

However, out of date is not a problem with mediator approach. This approach will provide various data from various sources in real time format. But this approach not very popular maybe because of biologist's tradition. Biologists have started using logically warehouse long before virtual database become evolve.

At this point, the ideal system as mention in the third paragraph will be achieved if the both approaches can be combined. As conclusion the aim of this

project is to study and analyze this suggested approach to find the best solution. The analyzing will be base on response time of retrieval data.

## 5.10 Statement of the Problem

From the problem background, the aim of this project is to integrate several biological databases using combination of warehousing and mediation approaches. The questions need to be answered in this project is, how can the mediator and warehousing approaches be combined and does the combination of mediated-warehousing approach have better presentation for heterogeneous biological data integration and provide acceptable performance in terms of response time.

## 5.11 Project Objective

The main goal of this project is to introduce mediated-warehousing approach in heterogeneous biological data integration.

The objectives of this project that need to be fulfilled to achieve this goal are:

i) To study and analyze strengthens and limitations of mediation and warehousing approaches in heterogeneous biological data integration application.
ii) To compare and analyze the response time of each query between mediator and data warehouse approaches.

## 5.12 Scope

The scopes of this project including:

i) This project focuses on combination of mediation and warehousing approaches in heterogeneous biological data integration only.

ii) This project will only consider on the response time of retrieval data for mediation, warehousing and mediated-warehousing approach.

iii) This project will cover on protein that is one of biological area and only three databases used as data sources.

## 5.13 Importance of Research Study

Nowadays, as reported in [5], the advent of the information superhighway has dramatically increased the need for efficient and flexible mechanism to provide integrated views over multiple information sources. The existence of various information sources in variety types and forms in such domain need to be have efficient information accessible and retrieval system.

Because of that necessity, information integration approach have been introduce to make several sources of information work together and provide useful and valuable information when it has been accessed from one-access-point. This approach can eliminate time constraint and in advance there is no dropped out information.

Currently, numbers of biological data have been increased and abundant of sources coming out without limitations. Phenomenon happens because people and environment keep changing and new technologies release base on experiences, experiments and research. Biologists and researchers need relevant and accurate information to help them in their hypothesis and experiments. Traditional biologists used their own data library to support their hypothesis. But in this recent year, they need to be able to share their data with others and at the same time avoided redundant work. Biological data integration is a one way how biologist can access data without to query each individual database.

However, single approach of data integration has own advantages and disadvantages. As example, warehousing as data materialization [6], where data from local sources are integrated into one single new database, on which all queries can operate. As stated in [7], one of miscellaneous of this approach is outdated information when that information is not maintained in the sources. Unlike warehousing, mediator approach supports virtual database [2] and transform query to sources, synthesize answer and then response to the user. As alternative to complement each advantages and disadvantages of both approaches, mediated-warehousing approach will be introduce.

### 5.13.1 Importance of Mediated-Warehousing Approach

Generally, importance of mediated-warehousing approach is to eliminate disadvantages of existing approach and keep in existence on the advantages. Firstly, this approach is important in term of representation novelty data to the user. Novelty data is important to help researchers and biologist in their experiments and to prove their hypothesis corresponding with time evolvement. Currently, only mediator approach serves this advantage and on the other hand data warehouse

needs regularly maintenance to present this kind of data and maintenance process needs a large amount of cost.

Next, this recommendation approach purpose is to reduce network bottleneck problem. Currently, network usage in transferring data is the most popular option and the number of data that available online growth rapidly. Mediator provides novelty data but not in term of time performance because if the query is complex, more time are need to present the result and sometime if the network performance is worst the result can't be send to the user. This limitation can be eliminating by materialized data as data warehouse approach.

Another issue is on cost in building and maintenance process, the aim of this suggestion approach is to reduce the cost. Using data warehouse approach, data must be moved or copied from exiting databases and data needs to be preprocessed into a common format. Data preprocessing includes data cleaning, integration and transformation and these tasks are very expensive.

Lastly, the important of this approach is to reduce semantic heterogeneity that will be supporting by ontology used. Mostly existing systems with mediator approach used domain ontology that serves as a global schema and plays a major role in resolving both semantic and syntactic variability.

## 5.14 Chapter Summary

This is introductory chapter that stated overview of the project consists of problem in heterogeneous biological data and the need of information integration.

General idea of this project is to analyze how mediated-warehousing approach can be applying in integrated heterogeneous of biological data sources. This chapter also presented the objectives that need to fulfill to achieve main goal of the project, scopes which means the boundary of the project and importance study of this project.

# REFERENCES


1. D. Page & M. Craven (2003). Biological Applications of Multi-Relational Data Mining. *SIGKDD Explorations,* 5(1):69-79.
2. Garcia-Molina, H., Ullman, J.D., &.Widom, J.D. Database Systems: The Complete Book. New Jersey: Prentice Hall. 2002.
3. Lambrix, P., & Jakoniene, V. (2003). Towards transparent access to multiple biological databanks. Paper presented at First Asia-Pacific Bioinformatics Conference (APBC2003), Adelaide, Australia.
4. Hernandez, T., & Subbarao Kambhampati. (2004). Integration of biological source: Current systems and challenges ahead. SIGMOD Record, 33 (3).
5. Hull, R., & Zhou, G. (1996). A framework for supporting data integration using the materialized and virtual approaches. In SIGMOD '96: Proceedings of the 1996 ACM SIGMOD international conference on Management of data at Montreal, Quebec, Canada, (pp.481-492). New York, NY, USA: ACM.
6. Ana Maria, C. M., Marcio, C. V., & Asterio, T. (2002). Combining Mediator and Data Warehouse Technologies for Developing Environmental Decision Support Systems. In M.J. Egenhofer and D.M. Mark (Eds.), GIScience 2002, (pp.196-208). Berlin: Springer-Verlag.
7. Widom, J. (1995). Research problem in data warehousing. Paper presented at Proceeding of 4th International Conference on Information and Knowledge Management (CIKM), Nov. 1995.
8. Jagadish, H.V., & Olken, F. (2003). Data management for the biosciences. Report of the NSF/NLM Workshop of Data Management for Molecular and Cell Biology, February 2-3, 2003. Available at http://www.eecs.umich.edu/~jag/wdmbio/wdmb_rpt.pdf.

9. Wooley, J.C., & Lin, H.S. (2005) Catalyzing Inquiry at the interface of computing & biology. Report of the National Research Council of the NationalAcademies. Available at http://genomicsgtl.energy.gov/pubs/NRCComputingandBiology/3OntheNatureofBiologicalData.pdf.
10. Karasavvas, K.A., Baldock, R, & Burger, A. (2004). Bioinformatics integration and agent technology. Journal of Biomedical Informatics, 37, 205-209. Available at www.sciencedirect.com.
11. Bitchutskiy, V., & Lathrop, R. Heterogeneous Biomedical Database Integration Using a Hybrid Strategy. Undergraduate research journal, University of California, Irvine.
12. Ng , A., Bursteinos, B., Gao, Q., Mollison, & E,.Zvelebil, M,. (2006). Resources for integrative systems biology: from data through databases to networks and dynamic system models. Briefings in Bioinformatics. Vol 7. No 4, 318-330. Oxford University.
13. Page, D., & Craven, M. Biological applications of multi-relational data mining. University of Wisconsin
14. Gene Expression Access on 1st May 2009 from http://genome.wellcome.ac.uk/doc_WTD020757.html
15. IntAct access on 1st May 2009 from http://www.ebi.ac.uk/Information/Brochures/pdf/IntAct08.pdf
16. BIND accessed on 1st May 2009 from http://nar.oxfordjournals.org/cgi/reprint/29/1/242.pdf
17. Wishart DS et al., *HMDB: the Human Metabolome Database*. *Nucleic Acids Res. 2007 Jan;35(Database issue):D521-6.*
18. RNAiDB accessed on 1st May 2009 from http://nematoda.bio.nyu.edu:8001/cgi-bin/about.cgi
19. National Human Genome Research Institute accessed on 1st May 2009 from http://www.genome.gov/27530687
20. The pathway resource list accessed on 15 April 2009 from http://www.pathguide.org/
21. Benson, D.A., Karsch_Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., & Wheeler, D.L. (2000) GenBank. Nucleic Acids Research, 2000, Vol. 28, No.

1, Oxford University Press. Available at http://nar.oxfordjournals.org/cgi/reprint/28/1/15.pdf

22. Sample GenBank Record (23 October 2006) accessed on 24th April 2009, from http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html.
23. Protein Data Bank (file format) (n.d) accessed on 24th April 2009, from http://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format)
24. Araujo, F.A., Pinheiro, A.M.A., Farias, K.M., Loscio, B.F., & Oliveira, D.M. (2008) FlagelLink: A decision support system for distributed flagellar data using data warehouse.
25. Kimball, R., & Caserta, J. The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data. Canada: Wiley Publishing. 2004.
26. Ji, Y. (2003) Towards framework for the virtual data warehouse.
27. Connolly, T.M., & Begg, C.E. Database systems: A practical approach to design, implementation, and management. Addison Wesley. 2002.
28. Lin H., Risch T., & Katchaounov T. (2000) Object-oriented mediator queries to XML data. Proceedings of the 1st International Conference on Web Information Systems Engineering (WISE2000), 2000; 38—45.
29. Brien, P. M., & Poulovassilis, A. (n.a) P2P query reformulation over Both-as-View data transformation rules.
30. Goble, C. A., Stevens, R., Ng, G., Bechhofer, S., Paton, N.W., Baker, P.G., Peim, M., & Brass, A. ( 2001) Transparent access to multiple bioinformatics information sources. IBM System Journal, Vol 40, No 2, 2001.
31. Lee, T.J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D. WJ., Tenenbaum, J.D., & Karp, P.D. (2006) BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics2006, 7:170.
32. Shah, S. P., Huang, Y., Xu, T., Yuen, M., Ling, J., & Ouellette, F. ( 2005). Atlas – a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005, 6:34.
33. Birkland, A., & Yona, G. (2006). BIOZON: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinformatics 2006, 7:70.

34. Rahm, E., Kirsten, T., & Lange, J. (2007). The GeWare data warehouse platform for the analysis of molecular-biological and clinical data. Journal of Integrative Bioinformatics .
35. Naidu, P. G., Palakal, M. J., & Hartanto, S. (2007). On-The-Fly data integration models for biological databases. SAC'07, March 11-15,2007, Seoul, Korea.
36. Mougin, F., Burgun, A., Bodenreider, O., Chabalier, J., Loreal, O., & Le Beux, P. (2008). Automatic methods for integrating biomedical data sources in a mediator-based system. In A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux (Eds.): DILS 2008, LNBI 5109, 2008 (pp.61-76). Berlin Heidelberg: Springer-Verlag
37. Lim, J.H., Park, I.S., & Hyun, S.J. (n.a) Wrapper-Mediator Based Web Source Integration of SNP Allele Frequency Data.
38. Smedley, D., Swertz, M.A., Wolstencroft, K., Practor, G., Zouberakis, M., Bard, J., Hancock, J.M., & Schofield, P. (2008). Solutions for data ntegration in functional genomics: a critical assessment and case study. Briefings In Bioinformatics, Vol. 9. (No 6.) 532-534.
39. Cadag, E., Louie, B., Myler, P.J., & Tarczy-Hornoch, P. (2007). Biomediator data integration and inference for functional annotation of anonymous sequences in Pacific Symposium on Biocomputing 12, (pp. 343-354). Seattle USA.
40. Connolly, T., & Begg, C. ( 2002) Database systems. (Third Edition). England : Addison-wesley.
41. Henniger, S., & Belkin, N.J. (n.d) Interface issues and interaction strategies for information retrieval systems. Access on May 27, 2009 from http://www.sigchi.org/chi96/proceedings/tutorial/Henninger/njb_txt.htm.
42. Honavar, V., Andorf, C., Caragea, D., Silvescu, A., Reinoso-Castillo, J., & Dobbs, D. (n.d) Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed, Autonomous Biological Data Sources. Artificial Intelligence Research Laboratory, Dept. of Computer Science, Iowa State University.
43. Lambrix, P., & Jakoniene, V. (2005). Ontology-based integration for bioinformatics. In proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005

44. Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hubner, S. (2001) Ontology-based Integration of Information-A Survey Existing Approaches. In Proceedings of IJCAI_01 Workshop: Ontologies and Information Sharing, Seattle, WA, 2001, Vol.pp.108-117.
45. Miled, Z., B., Li, N., & Bukhres, O. (2005) BACIIS: Biological and Chemical Information Integration System. Journal of Database Management, 16(3), 72-85, July-september 2005.
46. RCSB Protein Data Bank (PDB) accessed on 15 June from http://www.rcsb.org/pdb/home/home.do
47. Cerami, E. XML for Bioinformatics. New York: Springer. 2005.
48. Extensible Markup Language (XML) accessed on 1 July from http://www.w3.org/XML/.
49. ENZYME, Enzyme nomenclature database accessed on 1 July from http://www.expasy.ch/enzyme/
50. BRENDA, The comprehensive enzyme information system accessed on 1 July from http://www.brenda-enzymes.org/
51. Enzyme Nomenclature accessed on 1 July from http://www.chem.qmul.ac.uk/iubmb/enzyme/
52. Kuhlins, S., & Tredwell, R. (n.d) Toolkits for Generating Wrappers: A survey of software toolkits for automated data extraction from web sites. Department of Information Systems III, University of Mannheim, Germany.
53. IBM Informix Dynamic Server Performance Guide accessed on 5 August 2009 from http://publib.boulder.ibm.com/.
54. Ashish, N., A.Knoblock, C., & Shahabi, C. (n.d) Selectively Materializing Data in Mediators by Analyzing Source Structure, Query Distribution and Maintenance Cost. Information Sciences Institute, Integrated Media Systems Center and Department of Computer Science, University of Southern California.
55. Voisard, A. and Jürgens, M. 1999. Geospatial Information Extraction: Querying or Quarrying?. In Goodchild M, Egenhofer M, Fegeas R, Kottman C, eds. Interoperating Geographic Information Systems. 1st ed. Dordrecht: Kluwer Academic. 165–79.

56. Davidson S. B., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, C., & Stoeckert, C. (2000) K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. Center for Bioinformatics, Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia.