Intelligent Data Mining in Autonomous Heterogeneous Distributed and Dynamic Data Sources

Azra Shamim, Vimala Balakrishnan, Madiha Kazmi, and Zunaira Sattar

***Abstract*—**Data is increasing at a rapid speed. Data sources are often autonomous distributed heterogeneous and dynamic in nature. Changes in data sources should be propagated into data integration systems for valid, accurate and up to date decisions. In this paper, architecture of the intelligent data mining system for autonomous, heterogeneous, distributed and dynamic data sources is introduced. This architecture is enhanced version of intelligent data mining system in autonomous, heterogeneous and distributed bio database with data management, change detection and control layers. The enhanced architecture increase data organization, management and maintains updated data in the system.

***Keywords*—**Autonomous Distributed Heterogeneous and Dynamic Data Sources, Data Mining, Intelligent Data Mining, Expert System, Knowledge Base

I. Introduction

THE volume of data is increasing at a rapid pace. The amount of data stored in the biological databases has indeed grown exponentially over the past decade [1]. Data sets of interest to computational biologists are often heterogeneous in structure, content, and semantics [2]. Biological database are heterogeneous, distributed, and have different user interfaces [4]. Most of biological data is heterogeneous, distributed, autonomous and dynamic [3],[5]. Distributed database is a collection of multiple, logically interrelated databases distributed over a computer network [6], [7]. Distributed databases with no homogeneity in their data model and query language are called heterogeneous distributed databases [6]. Autonomy indicates the degree to which individual database can operate independently [7]. Autonomous heterogeneous distributed databases are the non-homogenous independent databases stored on multiple locations and linked together via a network [6].

The huge and growing amount of data is widely distributed over the internet in different online repositories. Biological resources are either publicly available on the Web, or local and private [8]. Biological data currently stored in biological databases, data warehouse (Bio Warehouse), flat files, relational, and object oriented databases. The term biological database is used loosely to refer a biological data collection in any of these forms [2],[6]. Bio warehouse is a biological data integration platform [9] ,[10].

The architecture of the intelligent data mining system for Autonomous Heterogeneous Distributed Bio Databases (discussed in section 2) has two problems. First, for valid and accurate decisions up to date data is required. However, the system does not propagate changes (updates) from dynamic data sources into the system to keep updated data. The system should include a mechanism to propagate changes into the system. Some mechanisms of change detection and control are discussed in [11],[12]. Authors integrate the mechanism discussed in [12] in this enhanced architecture. Second, there is no proper organization and management of the data in the system. Efficiency can be achieved by better organization and management of data. Furthermore, being a valuable asset; data must be protected from accidental loss. Cost of data lost is very high in a research environment. The system should be able to protect data from accidental loss or equipment failure.

In this paper, authors present an enhanced architecture of intelligent data mining system for autonomous heterogeneous distributed and dynamic data sources with data management, change indicator and controller layer. The proposed two layers overcome the problems of the system mentioned above. The rest of the paper is organized as follows: Section 2 presents literature review. The proposed architecture is described in Section 3. Section 4 discusses concluding remarks.

II. Literature Review

A. Data Mining

"*Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner*" [13]. "*The process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decision*" [14]. Data mining looks for interesting trends and useful patterns in data sets that reveal new insights for large data sets [23], [24], [25]. The architecture of data mining system is presented by Han et al. in [15]. Data mining

Azra Shamim is with FSKTM,University of Malaya, Kuala Lumpur Malaysia, (e-mail: azra.majeed864@yahoo.com).

Vimala Balakrishnan is with FSKTM,University of Malaya, Kuala Lumpur Malaysia, (e-mail: vimala.balakrishnan@um.edu.my).

Madiha Kazmi is with the National University of Science and Technology, Islamabad, Pakistan ; e-mail: madihakazmi@yahoo.com).

Zunaira Sattar, is with Computer Science Department, GC University Faisalabad, Pakistan (e-mail: zunaira_pk@yahoo.com).

is the process through which information that is actionable and valid is extracted from large databases [16].

B. Intelligent Data Mining System

Intelligent data mining is calling knowledge intelligently [6], [17] or intelligent query answering [6], [18]. The structure of intelligent data mining system based on the expert knowledge is presented by Junhua Hu et al. in [19]. Intelligent data mining system based on the expert knowledge follows the problem solving skeleton discussed by Takahara et al. in [20]. The architecture of intelligent data mining for Autonomous Heterogeneous and Distributed Bio Databases is discussed in [6]. This architecture consists of three layers; Data Extraction Layer, Data Mining Engine Layer and Expert System Layer. The preprocessing /preparation of data is performed by the Data Extraction Layer. This layer locates and access relevant data from autonomous heterogeneous distributed bio databases, clean, transform, optimize and load data in to the system. Data Mining Engines Layer performs a variety of data mining task on data. Expert system layer interacts with user to get a user query, analyze it, and select appropriate data mining technique/algorithm to be performed on data. Expert system provides the intelligence assistant to users.

C. Expert System

"*Expert system is a computer application that employs a set of rules based on human knowledge to solve problems that require human expertise"* [21]. Expert System is a computer system that applies reasoning methodologies to knowledge in a specific domain to render advice or recommendations much like a human expert [22].

III. Enhanced Architecture

The architecture of intelligent data mining for autonomous heterogeneous distributed and dynamic data sources is presented as layered approach as shown in Fig. 1. It consists of 7 layers; each layer is explained below:

A. Data Source Layer

This layer consists of heterogeneous, physical distributed, autonomously maintained and dynamic data sources. Data sources can be online public, private or commercial databases, bio warehouse and flat files.

B. Change Indicator and Control layer

This layer detects and propagates changes into the data mining mart layer. It insures the data consistency between data sources and data stored in the system.

C. Data Preprocessing Layer

This layer extracts data from autonomous heterogeneous distributed and dynamic data sources. After data extraction; preprocessing is performed by the layer. Data preprocessing include data cleansing, selection, transformation and loading into the data mining mart layer. The working of data preprocessing layer is controlled by expert system layer.

D. Data Mining Mart Layer

Data mining mart layer is a repository of relevant and valuable data. It contains data on which data mining task is to be performed by the data mining engine for search of useful, valid and actionable information. Three types of data are stored in the data mining mart; (i) local, (ii) back up and (iii) extracted.

i. Local Data

Data that is internal to an organization is called local data. Local data contains proprietary or experimental data to compare it with extracted data for decision making.

ii. Extracted Data

Data that extorted from autonomous, heterogeneous distributed and dynamic data sources is called extracted data.

iii. Back Up Data

This data is used for recovery purpose in case of accidental deletion or equipment failure.

E. Data Management Layer

The function of this layer is to organize data in to different clusters. It also informs the expert system layer about data organization and protects data from accidental loss or deletion or equipment failure. Data management layer has following components:

i. Clustering Engine

Clustering engine partitioned extracted data into heterogeneous clusters with homogeneous objects. Organization of data into clusters may provide high efficiency, reduced search space and response time. It has following components.

ii. Back Up Manager

Back up manager is responsible for protecting data from accidental deletion, lose and equipment failure. It is responsible for periodic data backups and works according to back up policy specified by the expert system layer. It stores data related to back up in the Meta data repository.

iii. Meta Data Manager

Meta data manager provides a way of communication between back up manager, Meta data repository and clustering engine. Clustering engine stores the Meta data about data organization and back up manager stores data related to backups with the help of the Meta data manager in Meta data repository.

iv. Meta Data Repository

Clustering engine creates the Meta data about data and stores it in the Meta data repository e.g. like number of objects in cluster and total number of clusters. It also contains data about the backups.

F. Data Mining Engine Layer

The job of data analysis is performed by the data mining engine layer. It performs the task of data mining in order to uncover valuable, interesting, and hidden information from data stored in the data mining mart. It has following components:

Data Mining Engine Controller: It gets instruction from the expert system layer. These instructions specify data mining technique to be performed and on which data cluster. Accordingly, it forward control to specific component to

performed specified technique. It also gets the result from specific component and forwards it to expert system layer for further evaluation.

Pattern Analyzer: It identifies different pattern in underlying data.

Data Classifier: It infers the class of a data item.

Association Analyzer: It identifies association between data items.

Cluster Analysis: It partitioned data set in to different clusters.

Network Modeler: It identifies and model biological networks.

G.Expert System Layer

It is the most important and intelligent layer of the system. It acts as a control unit that controls working of other layers of the system. It provides the expert/intelligent assistance to the users in response to their queries. It takes users query and provides answer of it to users. Expert system layer instructs data mining engine what to do and how to do? It gets high-level user query from users, transforms it into a low-level query. After converting query into low level; expert system layer interprets it and selects appropriate data mining technique required to answer the query. Selection of a data mining technique depends on nature of the user query. Once the selection of technique is done, expert system layer specifies the data mining engine which technique needed to be performed and on which dataset. Expert system layer obtains result from the data mining engine and then uses expert knowledge to evaluate and analyze the result. Evaluation and analysis process converts results of data mining engine into a real knowledge.

IV. Conclusion

There has been a dramatically boost in data generation in last few years. Not only data is increasing but its complexity too. Data is often heterogeneous, distributed, autonomous and dynamic in nature. Due to distribution, heterogeneity, volume and complexity data analysis, organization and management become difficult. Changes occurred in data sources should be propagated in to data integration systems to provide users up to data information. Proper data management and organization with in systems is necessary for better efficiency. In this paper, authors proposed the architecture of intelligent data mining system, which may help researchers in evaluation and analysis of heterogeneous, physical distributed, autonomously maintained and dynamic data. It collects data from multiple dispersed independent diverse and dynamic data sources and provides integrated view of data. Then extract valid, relevant, correct, and actionable information from data as well as knowledge. This mined knowledge is combined with expert knowledge to help decision making process.

References

[1] Thomas Hernandez, Subbarao Kambhampati, "Integration of Biological Sources: Current Systems and Challenges Ahead", SIGMOD Record, Vol. 33, No. 3, September 2004 http://dx.doi.org/10.1145/1031570.1031583

[2] Adrian Silvescu, Jaime Reinoso Castillo, Vasant Honavar, "Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed, Autonomous Biological Data Sources", In Proceedings of the IJCAI 2001 workshop on Knowledge Discovery form Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources, 2001

[3] Fasman, K., "Restructuring the Genome Data Base: A model for a federation of biological databases", Journal of Computational Biology, Volume 1, Number 2, pp. 165-171 http://dx.doi.org/10.1089/cmb.1994.1.165

[4] Iskandar Ishak, Naomie Salim, "Database Integration Approaches for Heterogeneous Biological Data Sources: An overview", Proceedings of the Postgraduate Annual Research Seminar 2006

[5] Doina Caragea, Jyotishman Pathak, Jie Bao, Adrian Silvescu, Carson Andorf, Drena Dobbs, Vasant Honavar, "Information Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data Sources", B. Lud¨ascher and L. Raschid (Eds.): Springer-Verlag Berlin Heidelberg, pp. 175–190, 2005.

[6] Azra Shamim, Maqbool U. Shaikh and Saif U. Rehman. "Intelligent Data Mining in Autonomous Hetereogeneous Distributed Bio-Databases". In proceedings of Second International IEEE Conference on Computer Engineering and Aplications (ICCEA-2010), Vol. 1, Pp: 6-10, March 19-21, 2010, Bali Indonesia.

[7] M. Tamer Ozsu, Patric Valduriez. (2003), Pricnciple of Distributed Database Systems, 2nd Ed.

[8] Zoe Lacroix , Omar Boucelma, Mehdi Essid, "The Biological Integration System", WIDM '03: Proceedings of the 5th ACM international workshop on Web information and data management, November 7–8, 2003, New Orleans, Louisiana, USA.

[9] Cao SL, Qin L, Wang W, Zhu YY, Li YX., "Applications of gene ontology in bio-data warehouse", In Proc Sixth Annual Bio-Ontologies Meeting, Brisbane, Australia, 2003: pp 33–36

[10] Cao SL, Li R, Zhang ZP, Zhu YY, Li YX., "BioDW: A platform of the integrated bioinformatics data warehouse", Computer Science, 2003, Volume 30, Number 10.B, 104–106

[11] Lean Yu, Wei Huang, Shouyang Wang, Kin Keung Lai , Web warehouse – a new web information fusion tool for web mining, Elsevier, information fusion, science direct 2006

[12] Saif Ur Rehman Malik, Maqbool Uddin Shaikh. "Web Warehouse: Towards Efficient Distributed Business Management", In proceedings of IEEE International Multi-Topic Conference 2008 (INMIC-2008), Karachi, Pakistan

[13] David Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining- MIT Press, Cambridge, MA, 2001.

[14] Thomas connolly ,Carolyn begg, Database Systems: A Practical Approach to Design, Implementation and Management .4th Edition, Addison-wesley, 2003

[15] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, New York: Morgan Kaufmann Publishers, 2006

[16] Hameed Hussain, Maqbool U. Shaikh and Saif U. Rehman. "Proposed Text Mining Framework to Explore Issues from Text in a Certain Domain". In proceedings of Second International IEEE Conference on Computer Engineering and Applications (ICCEA-2010), Vol. 1, pp: 16-21, March 19-21, 2010, Bali Indonesia.

[17] Jian Liang, Xiao Li, Hongshuo Liu et al. (2002).'The Design and Realization of Intellectualized Data Mining System' In: Application Research of Computer, pp.89-91.

[18] Chen Yi-ming. (2002).'Data mining technique and intelligent query answering associated analyzing' In: Journal of Northwest Normal University (Natural Science). (38) pp.41-43, pp.67.

[19] Junhua Hu, Yongmei Liu. (2006). 'Designing and Realization of Intelligent Data Mining System Based on Expert Knowledge': IEEE International Conference on Management of Innovation and Technology June .2006. pp. 380 – 383

[20] Y. Takahara, Y. Liu, J. Hu et al, "Intelligent Data Mining System", Proceedings of International Conference on E-Business, Beijing, 2002, pp 274-280

[21] George M. Marakas, Decision Support Systems in the 21st Century, 2nd Edition, Prentice Hall, 2002

[22] Efraim Turban, Jay E. Aronson, Narasimha Bolloju, Decision Support Systems and Intelligent Systems, 7th edition, Prentice Hall College Div, 2001

[23] Ann Okerson, Text & Data Mining - A Librarian Overview, IFLA WLIC 2013.

[24] Tipawan Silwattananusarn1 and KulthidaTuamsuk, Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012

[25] Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining, Report from the Expert Group, 2014

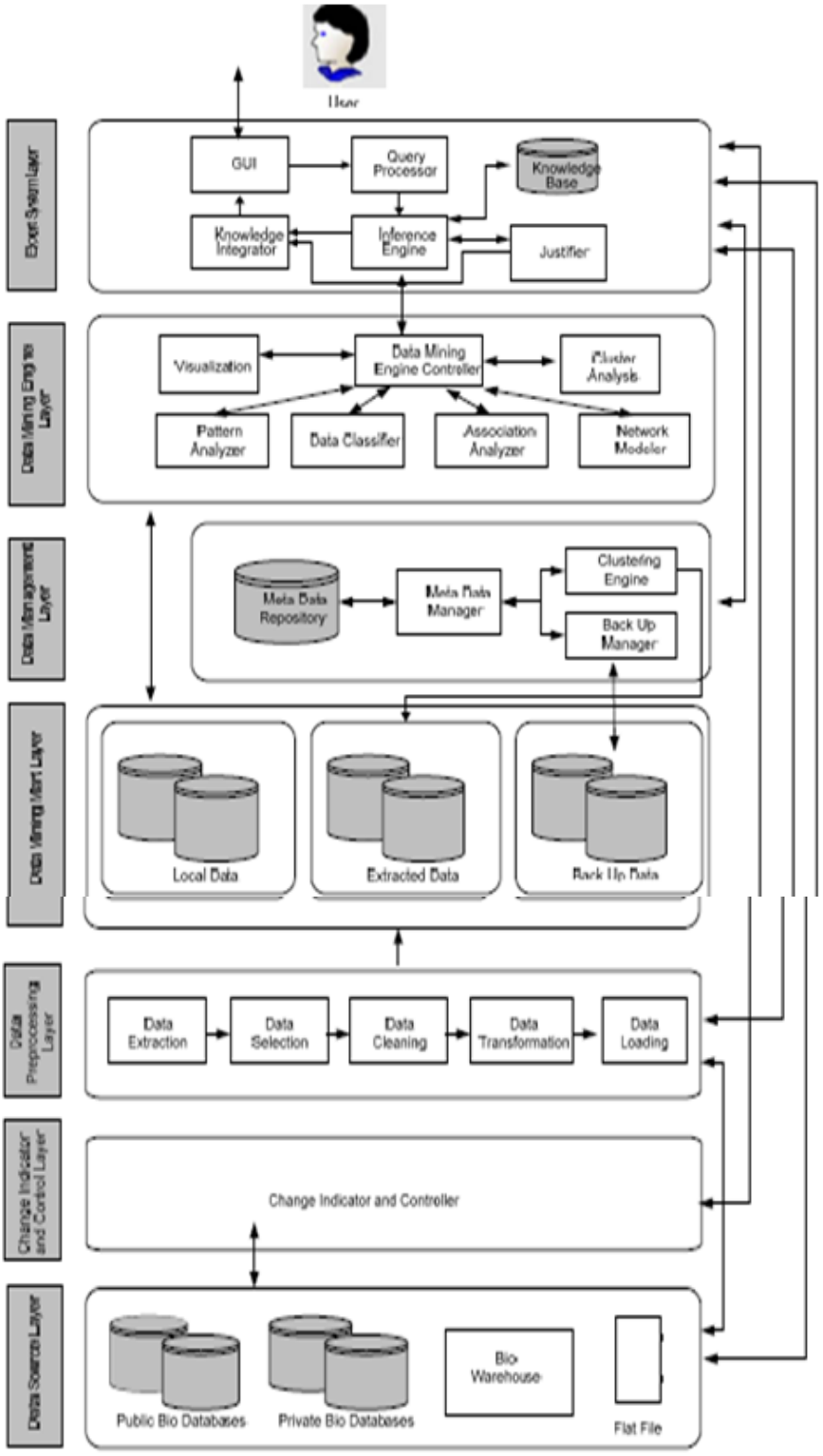

Fig. 1 Architecture of Intelligent System Data Mining in Autonomous Heterogeneous Distributed and Dynamic Data Sources