DEALING WITH HETEROGENEITY OF DATA AND KNOWLEDGE USING OBJECT MODELING

Sarswati Kumar Pandey and Madhulika
[1]Department of Computer Science and Engineering,
Amity University, Noida, Uttar Pradesh, India

ABSTRACT
Integration of a large amount of heterogeneity of data and knowledge sources is the biggest challenge in Knowledge representation. In many cases while integrating different types of data, objects and knowledge sources is larger than the sum of its parts. We are integrating data to gain valuable insights produced by different scientific experiments. In this research we examine some of the issues in semantic heterogeneity and propose a novel architecture for resolving such problems. The approach involves the use of Artificial Intelligence tools and techniques to construct "object oriented models," that is data and knowledge representations of the constituent databases and an overall domain model of the semantic interactions among the databases. These domain models are represented as Knowledge Sources (KSs) in blackboard architecture. This architecture lends itself to an opportunistic approach to query processing and goal-directed problem solving. We introduce the notion of Data/Knowledge packets as a means of supporting both operational and structural semantic heterogeneity. Using object modeling approach we are able to integrate the data and knowledge which will not be redundant and much complex. It will improve the decision making techniques in business domains.

KEYWORDS: *Heterogeneity, Integration, Object Modeling, Data and Knowledge*

I. INTRODUCTION

Dealing with heterogeneity of data and knowledge is one of the biggest challenge and opportunities for knowledge representation. The challenges are in integrating the heterogeneous data and knowledge from different data sources such as DBpedia, Wikipedia and Google etc. Heterogeneity is of different types-

- Heterogeneity of vocabularies, accuracy, level of abstraction either low level of abstraction or high level of abstraction.
- Heterogeneity of data and information structure, syntax, semantics.
- Data items having different Ids for the same data objects.
- Data acquired from rapidly reproduction via the shared medium on the World Wide Web.

In many scenarios, while integrating the different objects, data, information and knowledge sources results in much larger than that of the sum of its parts collected from different data and knowledge sources. The internet era has increased the production of data and information in digital format. The people has an easiest way to access the data and knowledge which was not possible much before and the rapid increase in digital data has some major disadvantages like duplicity, inconsistency and rapid transmission of all these digital data across the Net. The information is widely spreaded across the World Wide Web and it is more difficult to make decision using the right data. Now digital information age is having a critical issue that how the knowledge is being processes behind the large amount of information that reflects every day through every aspects of normal life like news, social networking, social media, radio, TV, e-mails, blogs, research papers so on. Everywhere the quantity of information is exponentially increasing and the quality of information is rapidly decreasing. We can have much valuable information by integrating the data and knowledge produce by different experiments made by the scientists and researchers. Researchers work on different techniques of

integrating heterogeneous data and knowledge from tight coupling of the task requires. Solving the problem of heterogeneity of data and knowledge still has many challenges. Data Mining is also used to find patterns from huge databases [11]. To cope in the future with those challenges needs some advancement in the integration of data and knowledge.

- Many government programs are allowing mandating the sharing of data which results in more collaborative and integrative opportunities.
- Some systems such as freebase, DBpedia (Wikipedia) and Wikidata allow people to integrate their own data, information, facts and figures into sharable environment such as cloud.

Rapidly increase in cheap storage, parallel hardware and parallel software advances the integration of data and knowledge form different knowledge sources.

II. RELATED WORK

To deal with integration of heterogeneous data and knowledge object modeling approach had been used to integrate and omit the duplicity and retransmission of data and information on a sharable and synchronized medium.

Large heterogeneous data: Linked Open Data

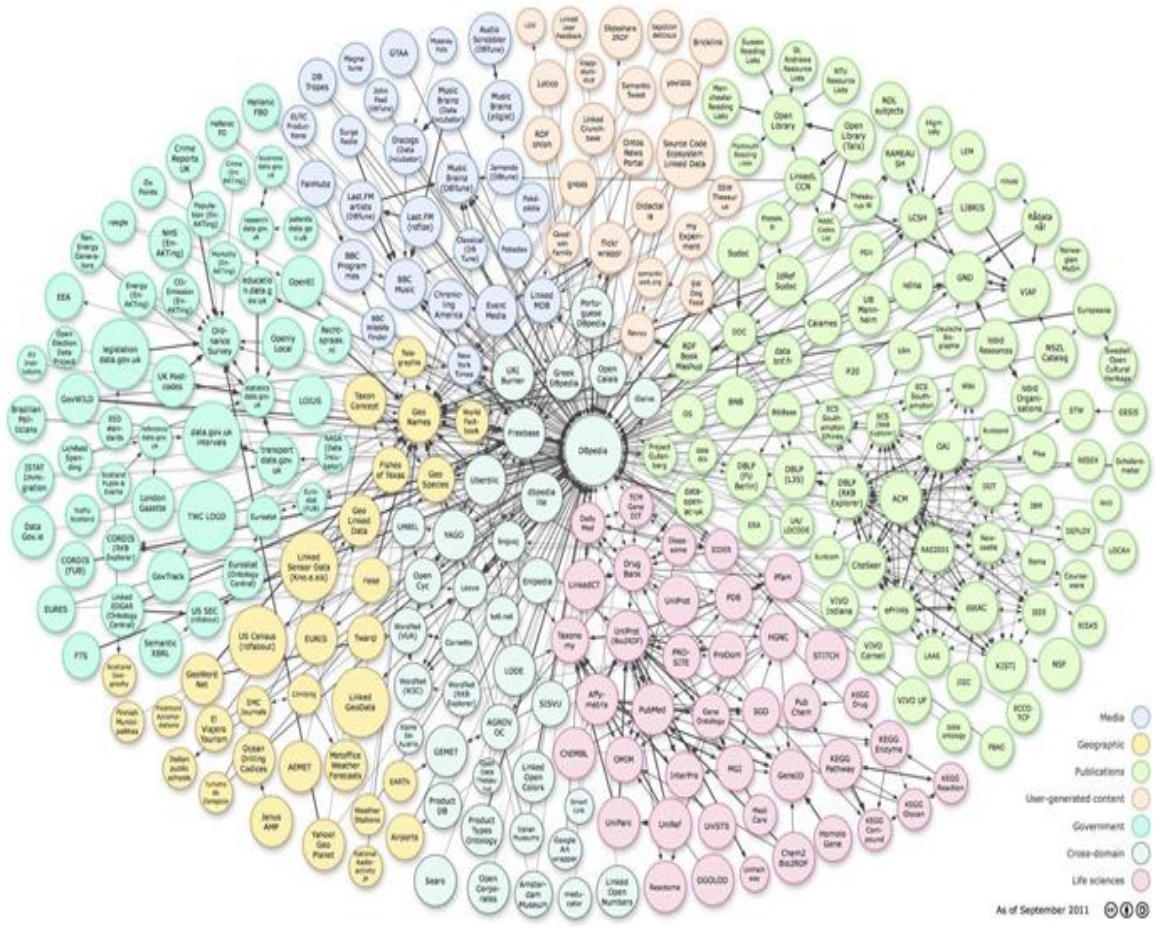

Fig. 1 Linked Open Data [1]

At the begin of the 21st century it was estimated that the total unique information produced in the world was one to two Exabyte every year. After three years the same authors updated their facts and they estimated that this quantity had increased between 3 to 5 Exabyte per year. The estimation in 2005 by some of the Economists was that the mankind created 150 Exabyte of data and in the year 2010 it was 1200 Exabyte that means a lot of data in the form of texts, numbers, images, sounds, etc. All these data are much important for the human beings for different purposes [5].

The digital revolution has so many advantages involving access to the knowledge and dissemination due to in the knowledge economy information is becoming the most valuable commodity especially

when an open access strategy is presented but it also suffers with the problems like the capacity to the process, understanding and taking advantage from this huge world information production and accumulation conformed as digital or electronics documents, which is also referred as information overload. The information overload may be defined as like a person felt too much difficulties while understanding the information and making a proper decision that caused by the presence of a lot of huge information. Information overload in the digital context is caused by different issues as rapidly increasing rate of new data, and due to this the duplicity of the data on the internet [5]. Individuals are affected by this information overload being distracted in their work environments and being confused in their personal decision making and companies are also using the same bad strategies in their information management that leads them to poorer decision making. This information overload and decision making process can be reduced by using the new technological advancements like Business Intelligent System based on Soft computing techniques, data mining strategies, visualization techniques and human-computer interaction etc., may support better more effectively and efficiently the decision making processes[9]. However the decision making processes are based on human process as well as computer process using artificial intelligent agents based systems.

To convey the information using visualization techniques at the high bandwidth of the perceptual system of the human requires recognition of the patterns and supporting navigations. Visual Analytics integrates the advantages of machines with the human strengths like intuition, problem solving, visual perception, analysis. Thus the key components of the knowledge discovery are patterns of knowledge, data and knowledge of domain. There is a database known as federated database which map different database into single database and these databases are connected with the help of computer network and many network management protocols manage these databases [10]. NFS Network File System) provide users with a unified view of data stored on different file systems [12].

III. OPTIMIZED OBJECT ORIENTED MODELING - A METHODOLOGY

Optimized object oriented modeling includes the objects retrieved from different data sources, projects the columns based on the attributes from the different data sources according to user's need and take the union of all projected data sources and if there are some common attributes in any other data set then project that data source's attributes then take the cross products with the union of all projected data sources.

Query Interface

Query Interface includes two main components. The clients interact with the query interface. The clients enter the plain text and plain text is being formulated as query by using the Query Formulation. Result Representation is used to show the fetched result from different data sources based on client's query condition. Sometimes result needs to be decomposed so the result goes to answer decomposition components and the result is decomposed into normalized forms. When the clients enter the query in de-normalized form, the query is first send to the query decomposition component and query is decomposed using the various normalization techniques [3].

Data Modeling Engine

Data modeling engine has one important component: The query processor. The query processor is used for projection, selection, union and cross multiplication. When the query comes to data modeling engine the query is sent to different data sources. Data is stored in multiple formats in different data sources. The data is projected, selected according to the user's query and being sent to the query modeling engine. The result is sent to the query interface and the result may be represented as tabular format according to client's query.

Data Sources

Data sources are different central servers and storage where data is stored in de-normalized formats. The data sources stored structured, semi-structured and unstructured data. The heterogeneity of data and knowledge is the root cause of the weak decision making and redundant data.

Query Answering Engine

Query answering engine is responsible for dealing with the result and decomposition of the query and results. The query is decomposed when the clients submit the query in de-normalized format and query is sent to the query interface to the query modeling engine. The answer decomposition engine is

responsible for normalizing the results retrieved from the query modeling engine. The normalized data is sent to the query interface for representing the results to the clients.

Projection and Selection

Projection and Selection both is relational operator used in the database. Selection is the process of taking the horizontal rows subset of a table that must be satisfying some conditions based on user's query. Select statement returns the combination of rows and columns in a table. Selection is implemented to the where clause of a SELECT statement. Projection is implemented through the projection operators. Projection is defined as taking the vertical subsets of the columns of a single table that outputs the unique rows.

Joining

A join is used among two or more tables that are connected through the common columns that create a new result table [4].

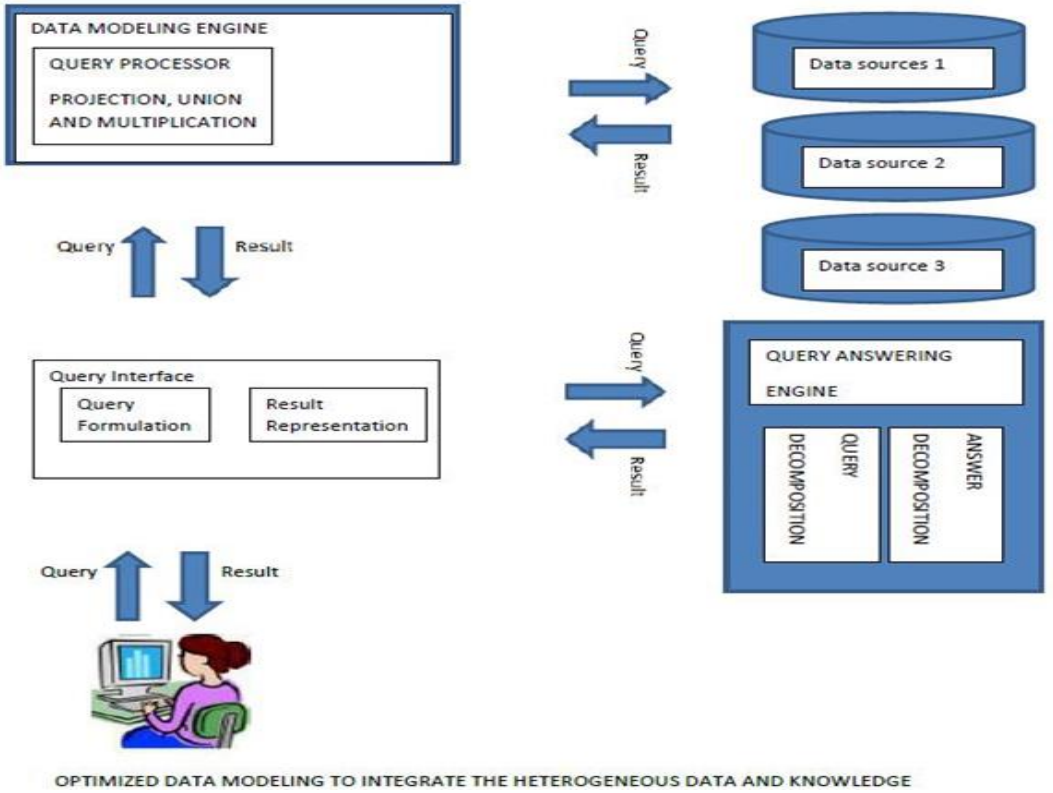

Fig. 2 Optimized Object Oriented Modeling to Integrate Heterogeneous Data and Knowledge

Illustrative Examples

Let's consider a user wants An Institution's M. Tech. student's record (enroll_ id, name). The user enters a query in an optimized object modeling application using query interface. The query is formulated using query formulation. Data Modeling Engine sends the query to the data sources and data sources results in three different results for M. Tech. as section1, section2, section3 with different attributes. Data source1 results with attributes enroll_ id, name, phone_ no, and data source2 results with attributes enroll_ id, name, phone_ no, e- mail, address and data source3 results in enroll_id, name, e- mail. In Data modeling engine Query Processor projects the enroll_ id, name from all the results and tales union of all the projected tables. The query processor sends the result to the query interface. In query interface the result representation results to the user's query in well formatted manner. If there is need of query decomposition and/or result decomposition then the query interface sends the query to the Query Answering Engine. Query Answering Engine decomposes the query and/or answer and results to the query interface.

IV. CONCLUSIONS

To deal with heterogeneity of data and knowledge is still a biggest challenge. Using optimized object oriented modeling this challenge can be reduced till some extent but still can not be solved completely. This approach mostly focuses on structured objects and semantic web data. Using object modeling integration becomes consistent and easier to understand and to take decision more accurately. It improves the decision making in business analysis. Object oriented modelling automate heterogeneous knowledge systems evolution & mapping maintenance. It maximizes the storage performance and reduces memory and processing costs.

V. FUTURE WORKS

In future the integration of heterogeneous data and knowledge may be fully solved using cloud computing, hybrid object oriented modeling and big data analytics. Object oriented modelling can be used in the distributed environment and can be used with the parallel systems to optimized the execution cost and increase the performance of the model to integrate the heterogeneity of various data sources.

ACKNOWLEDGEMENTS

The author would like to thank Ms. Madhulika her involvement and support in the completion of this paper.

REFERENCES

[1] Noy, Natasha, Deborah McGuinness, and Yuliya Lierler. "Research Challenges and Opportunities in Knowledge Representation, Section 2.4. 2 Advances in satisfiability and answer set programming." (2013).
[2] Josifovski, Vanja, and Tore Risch. "Functional query optimization over object-oriented views for data integration." *Journal of Intelligent Information Systems*12.2-3 (1999): 165-190.
[3] Risch, Tore, and Vanja Josifovski. "Distributed data integration by object-oriented mediator servers." *Concurrency and computation: Practice and experience* 13.11 (2001): 933-953.
[4] Caragea, Doina, et al. "Information integration and knowledge acquisition from semantically heterogeneous biological data sources." *Data Integration in the Life Sciences*. Springer Berlin Heidelberg, 2005.
[5] García Peñalvo, Francisco José, Ricardo Colomo Palacios, and Jane Yung-Jen Hsu. "Discovering knowledge through highly interactive information based systems." (2013).
[6] Pedrinaci, Carlos, and John Domingue. "Toward the Next Wave of Services: Linked Services for the Web of Data." *J. UCS* 16.13 (2010): 1694-1719.
[7] Kerschberg, Larry. "Knowledge management in heterogeneous data warehouse environments." *Data Warehousing and Knowledge Discovery*. Springer Berlin Heidelberg, 2001. 1-10.
[8] Sun, Yizhou, et al. "Mining knowledge from interconnected data: a heterogeneous information network analysis approach." *Proceedings of the VLDB Endowment* 5.12 (2012): 2022-2023.
[9] Madhurima, Madhulika. "Object tracking in a video sequence using Mean-Shift Based Approach: An Implementation using MATLAB7." *IJCEM International Journal of Computational Engineering & Management* 11 (2011).
[10] Bhatia, Madhulika. Introduction to computer network. Madhulika, 2009.
[11] SRIVASTAVA, SMRITI, and ANCHAL GARG. "DATA MINING FOR CREDIT CARD RISK ANALYSIS: A REVIEW." International Journal of Computer Science (2013).
[12] Hooda, Madhurima, and Madhulika Bhadauria. "Recent innovations in Distributed Systems: Challenges and Benefits." *INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY* 10.10
[13] (2013): 2057-2061.

AUTHORS

Sarswati Kumar Pandey is a student of M. Tech Computer Science & Engineering in Amity School of Engineering and Technology [2013- 2015], Amity University. He has done B.Tech in Computer Science & Engineering [2008-2012] from Bhabha College of Engineering affiliated

with Gautam Buddha Technical University. He is currently working on research area in BIG DATA, Knowledge Representation and Artificial Intelligence. He attended workshop in "I-Care 2013 5th IBM Collaborative Academia Research Exchange" October 17-19, 2013 organized at IBM Research Lab, New Delhi and also participated in "CONFLUENCE 2013 - The Next Generation Information Technology Summit" in the 4th International Conference at Amity University September 26-27, 2013. He has done Advance training for Professionals Sponsored by "Department of Electronics and Information Technology" Govt. of India, New Delhi from 01st April 2013 to 30th June 2013 on C++, J2SE, J2EE, C# & ASP.Net, Android, Cloud Computing, Oracle and Soft skills at ABV-IIITM Gwalior. He has also done Power Searching with Google Online Programme by Google in October 2012.

Madhulika is working as an Assistant Professor in Department of Computer Science and Engineering at Amity University, Noida. She holds diploma in Computer Science Engineering, B.E in Computer Science Engineering, MBA in Information Technology, M.Tech in Computer Science & Pursuing Ph.D from Amity University, Noida. She has total 8 years of Teaching experience. She published almost 15 Research Papers in National, International conferences and journals.