Contingent Models for Integrating Multi-Sourced Heterogeneous Data

Aruna Animish Pavate
Assistant Professor
Department of Computer Engineering
Atharva College of Engineering, Mumbai University, Mumbai, India

Abstract— In recent years in addition to operational databases, web databases have become an important data source for most of the organization's. The web databases may be structured, semi-structured or unstructured such as files, HTML pages, etc. and contains several heterogeneity issues. Heterogeneity may be present in different forms, hence the information retrieval from heterogeneous data sources become a tedious task. As the number of sources increases, need more intelligent retrieval techniques. There are number of techniques are available to remove the heterogeneity but the main focus is on information content and semantics of the data. The use of semantic web technologies such as ontology helps to represent such web data into clear information which can be incorporated in a data Warehouse. Ontology also helps to represent the industry requirements in a prescribed way, which needs to be used during the design. This report presents an ontology driven methods which helps to integrate data sources and business requirements from heterogeneous sources. Currently ontologies are being commonly used for dealing semantic knowledge, Information retrieval from distributed heterogeneous data sources. This survey describes different systems and comparisons between them for ontology -based data integration. This paper shows similarities and differences among the systems by comparison and classification.

Key words: Web databases; Heterogeneous; Semantic Web; Ontology; Data Integration

I. INTRODUCTION

Many applications data or information fusion, data mining and decision aids need to access multiple heterogeneous data sources. These data sources may come from internal and external databases. Any change in application domain induces semantic change in the data sources. For these applications data/information coming from different sources must be corrected, aligned, collected, combined and aggregated. In order to make collaborative work between applications multiple data sources, must be linked and integrated to be processed. Integrating data from heterogeneous sources and retrieval is a not an easy task. As now a day's information is distributed among different locations and stored in different format. Data integration is related with merging data that share some common semantics but originate from unrelated sources. Because many organizations put in information on distributed databases, they prerequisite a way to retrieve data from different sources and assemble it in a unified way.

In reality, data integration is a complicated discipline. There isn't a universal approach to data integration, and many of the methods IT experts use are still developing. Some data integration approaches might work better than others on behalf of an organization, depending upon that organization's requirements. Necessarily, when we work on data integration, we must takeinto consideration a more significant and compound concept called "heterogeneity". Heterogeneity may be present in from of networks, computer hardware, operating systems, programming languages, implementations of different developers. Heterogeneity might be categorized into four such as:Structural heterogeneity, Syntactical heterogeneity, Systemic heterogeneity, Semantics heterogeneity.

There are several methods created to address the problem of dealing with different constructs and interpretations. In general, the methodologies might be distributed into two different branches: methodswith ontologies and methods without using ontologies (for example using metadata).

A. Data Integration:

Data integration involves merging data exist in different sources and providing users with a unified view of these data. Issues with combining heterogeneous data sources under a single query interface have existed for some time. There are many data integration technologies that serve the data acquisition needs of a data warehouse, and the demand for low-latentperiod data is causing IT organizations to evaluate a range of approaches: intraday batch extract, mini batches; enterprise application integration (EAI); extract, load, and transform (ELT) technologies; as well as real-time change data capture (CDC) techniques.

The challenge is to determine which solution or combination of solutions will meet the need for current data, which will propel the move to operational data warehousing.Data integration is involved with uniting data that share some common semantics but originate from unrelated sources. As of 2010 [update] particular of the work in data integration research involved the semantic integration problem. This problem reports not the structuring of the architecture of the integration, but how to resolve semantic conflicts between heterogeneous data sources. A collective strategy for the determination of such problemsinvolves the use of ontologies which explicitly define schema terms and thus help to resolve semantic conflicts. This approach represents ontology-based data integration. On the other way, the problem of relating research results from different bioinformatics repositories requires bench-marking of the similarities, computed from different data sources, on a single

criterion such as positive predictive value. This enables the data sources to be straight comparable and can be unify even when the natures of experiments are distinct.[3]

B. Data Integration Issues:

1) Data Heterogeneity (Structural, Syntactical, Systemic, Semantic):

When we work on data integration, we must consider a more important and complex concept called heterogeneity. Heterogeneity may be present in one of the above from i.e. structural (schema), syntactical (format), systemic (hardware and OS), semantic (meaning). Data may be stored in different structure, different format and may be using different platform and may have different meaning for the same terms or same meaning for the different terms. All these issues must be considered while integrating data from heterogeneous sources.

2) Constructing The Multidimensional Model Is Time Costing:

Computers can only present user information but they cannot understand the information well enough to display the data i.e. most relevant in a given circumstance. Web browsers, web servers or search engines forced to do keyword matching only, so need to reconstruct the internet so that computers not only present the information contained in the internet but also understand and make intelligent decisions on our behalf

3) Too Much Manual Work So More Automation Needed:

Example: Consider the case in which we needed to book an airline ticket. We want to have an automated agent that can help us to find the flight, query the price and finally book the ticket,however to automatically composite and invoke these application, thefirst step is to discover them, if you think about this process you will realize that almost all your manual work is spent on discovery of these service. The first step of integration is to find the components that need to be integrated in a more efficient manner. This process involves too much manual work.

4) Need To Discover Web Services:

The discovery of web services has specific requirements and challenges as compared to previous work on information retrieval systems and information integration systems. Several issues need to be considered:

1) Precision of the discovery process. The search has to be based, not only on syntactic information, but also on data, functional and QoS semantics.
2) Enable the automatic determination of the degree of integration of the discovered web services and a web process host.
3) The integration and interoperation of web services differs from previous work on schema integration due to the polarity of the schema that must be integrated.

C. Ontology:

In computer science and information science, ontology formally represents knowledge as a hierarchy of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts. [1][2] Example: If two companies merge their databases, certain concepts and definitions in their respective schema like “earnings” have different meanings. In one database profits in dollars (a floating –point number) while in other it might represent the number of sales (an integer). Here a common strategy for the problem is the use of ontologies defined in term of schema which resolve problem. "Ontology is a shared conceptual model of clear formal specification." [5]

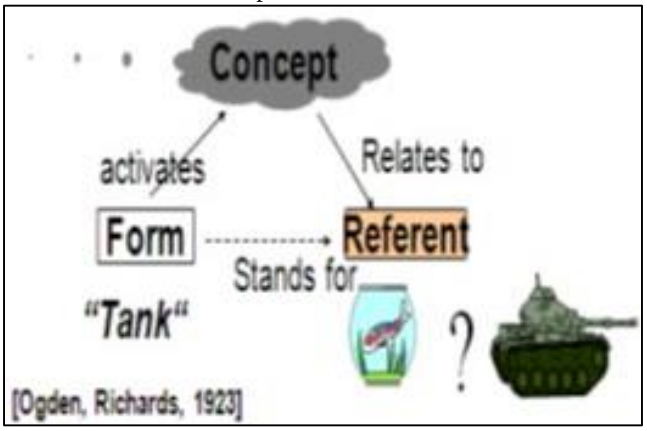

Fig. 1: Example of Ontology

Figure 1 represents example of ontology here if we consider tank then that can be referred as fish tank or tank which is to be used in war, certain concepts and definitions in their respective schema have different meanings. Here a common strategy for the problem is the use of ontologies defined in term of schema which resolve problem.

D. Advantages of Using Ontologies In Data Integration Process:

Three main causes for semantic heterogeneity:

1) “Confounding conflicts occur when information items seem to have the same meaning, but differ in reality e.g., due to different temporal contexts.
2) Scaling conflicts occur when different reference systems are used to measure a value. Examples are different currencies.

3) Naming conflicts occur when naming schemes of information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms"

Here focus is on the use of ontologies because of their advantages when using for data integration.[7] Among them:

- The vocabulary provided by the ontology serves as a stable conceptual interface to the databases and is independent of the database schemas,
- The language used by the ontology is expressive enough to address the complexity of queries typical of decision-support applications,
- Knowledge represented by the ontology is sufficiently comprehensive to support translation of all the relevant information sources into its common frame of reference, and
- The ontology supports consistent management and recognition of inconsistent data.

E. Scope:

Initially, ontologies are introduced as an "explicit specification of a conceptualization" [Gruber, 1993]. Therefore, ontologies can be used in an integration task to describe the semantics of the information sources and to make the contents explicit. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts. However, in several projects ontologies take over additional tasks. Some approaches use ontologies not only for content explication, but also either as a global query model or for the verification of the (user-defined or system-generated) integration description.

F. Objective:

To analyze the various methods of data integration based on ontology to remove the semantic heterogeneity of data while integrating information from various sources to get the richer semantics of data.

II. LITERATURE SURVEY

Many approaches are proposed to solve the data heterogeneity problem. Ontology method plays an important role by providing common definition and relationship of the concepts related to the business domain. The ontology approach is used in a number of enterprise practical applications. Following are the techniques which are used in integrating data from heterogeneous sources to remove semantic heterogeneity. Following are the techniques which have studied and make comparison between these techniques. We evaluate approaches according to four main criteria.

- Use of Ontologies: The role and the architecture of the ontologies influence heavily the representation formalism of ontology.
- Ontology Representation: Depending on the use of the ontology, the representation capabilities differ from approach to approach.
- Use of Mappings: In order to support the integration process the ontologies have to be linked to actual information. If several ontologies are used in an integration system, mapping between the ontologies are also important.
- Ontology Engineering: Before an integration of information sources can begin the appropriate ontologies have to be acquired or to be selected for reuse. How does the integration approach support the acquisition or reuse of ontologies?

A. Domain Ontology Metadata Model:

In [8] a Hybrid-based Recommender system driven by multi-ontology models is presented. The domain ontology is used in the context-based model to signify the learning materials. The learner's ontology model represents a subset of the domain ontology. The context-based model is combined with traditional rule-base model to provide the user with hybrid recommendations. : This model uses methods that combine the Content-based, Collaborative filtering, Rule-based to those that limiting particular conditions models, thus trying to avoid certain limitations in each one of the separate models. This method introduces a multi-model ontology-based framework for semantic search of educational content in ELearning repository of courses, lectures, multimedia resources etc.[5]

In the field of data integration, different kinds of method are proposed especially for enterprise level information system applications. In [10], author introduced a modeling language called GPM (Generic Product Model) proposed by Hitachi company to share data coming from several domain applications.

In this method, the domain ontology is embedded in the metadata of the data warehouse. Hence, the data record could be mapped from data bases to ontology classes of Web Ontology Language (OWL). As result, the accessing of information resources could be done more efficiently. The method is testing in a hospital data warehouse project, and the result shows that ontology method plays an important role in the process of data integration by providing common descriptions of the concepts and relationships of data items, and medical domain ontology in the ETL process is of practical feasibility.[5]

As shown in figure 2 the ontology approach to ETL needs to build domain ontology at first. The ontology is embedded in the metadata of the data warehouse to describe the semantic meanings of the properties of the database. When ETL meets different properties in data sources, it could map the record of the database to suitable fact tables according to the ontology concept definitions.

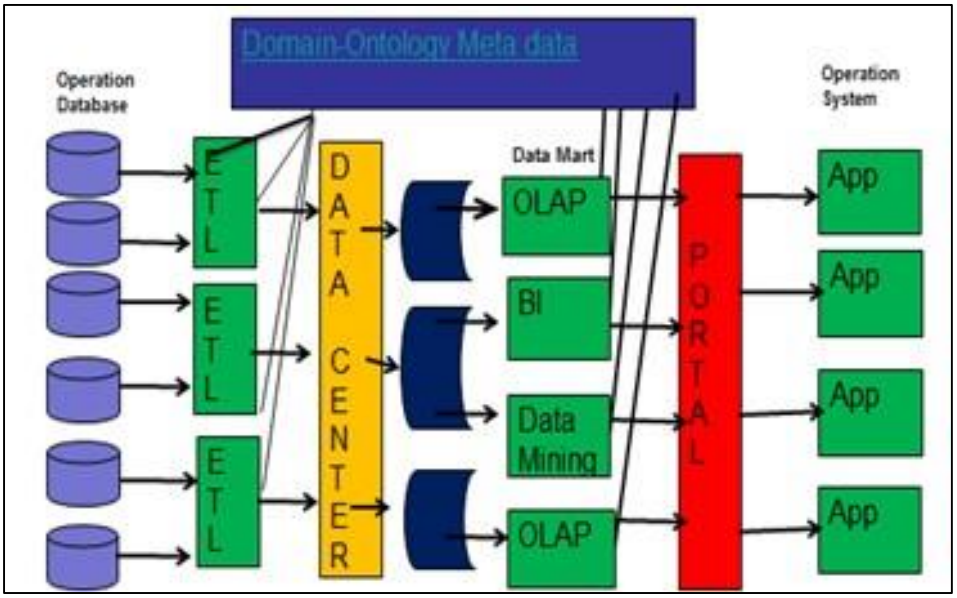

Fig. 2: Data ManagementArchitecture based ontology [1]

1) Advantages:

- The data record could be mapped from data bases to ontology classes of Web Ontology Language (OWL). As result, the accessing of information resources could be done more efficiently.
- This method focus on the use of ontology to data warehouse development which reduces the time cost of the ETL process and also optimizes the performance of the data warehouse.
- After the process of ETL, data records could be cleansed and stored along with the domain ontology. As a result, the data mining could be more flexible and efficient.
- Without the local ontology, the contribution of the ontology is simpler and the approach in our research is more suitable in the case of global ontology

2) Disadvantages:

- Although the study of ontology method to ETL and data warehouse grows fast, the problem of how to develop the domain ontology used in the ETL is not thoroughly researched.

B. Heterogeneous Data Sharing Method Based On Ontolog:

The heterogeneous data sharing and data query model [3]as shown in Figure.3. Model is divided into the physical layer, middle layer and application layer. By a direct interaction with the underlying physical layer is responsible for.

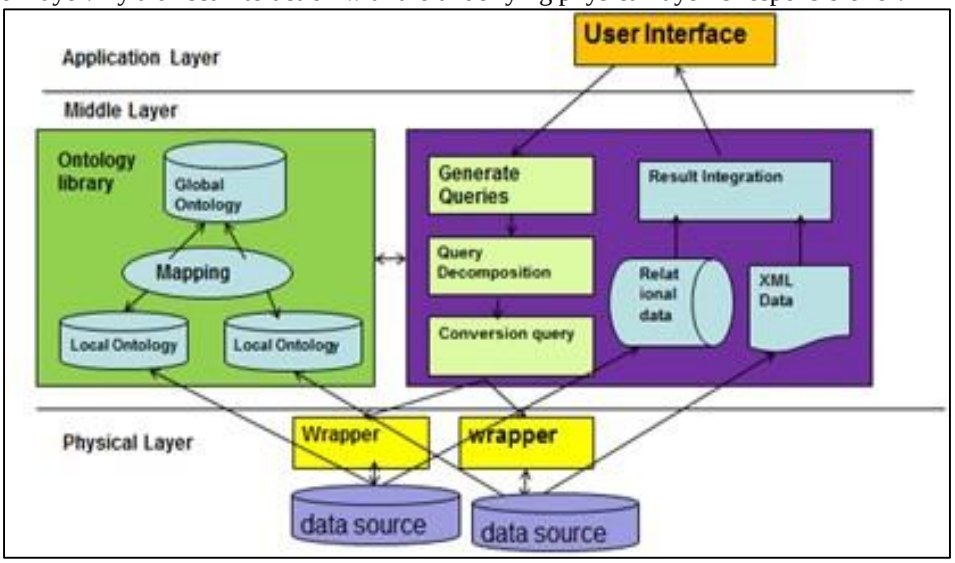

Fig. 3: heterogeneous data sharing and data query model [3]

Physical layer maps the logical model into various lands of storage formats for operational information. Middle layer is responsible for receiving user's request, the core of the system. Interact directly with the user application layer, is responsible to provide users with data query interface. User submits a query is based on the overall global ontology model, users do not care about the underlying data source, distribution and heterogeneity of data sources, query processing done by the middle layer processing.

1) Advantages:

- Combination local DW models of smaller scale to global DW model of larger scale through ontology merging process.
- Extract ontology from E-R schema of data sources to reduce the complexity of ontology construction.

C. Disadvantages:

- The method of recognizing facts and dimensions candidates from ontology system is also needed to be refined.

D. Local To Global Ontology Mapping:

An ontology-based data integration system [4] which consists in building a global ontology from the local ontologies corresponding to the data sources as opposed to a federated system approach. The role of the data integration system, which may be designed as a semantic portal for end users at the organization level is to exploit the global ontology and its integration with the local ontologies of data sources, as illustrated in Figure 4.

In this architecture, the data integration system constitutes a virtual database as opposed to a data warehouse, which copies data from several data sources in a single database. Now, the mediator maps the requests and answers between the global ontology/schema and the local ontologies with their associated source schemas.

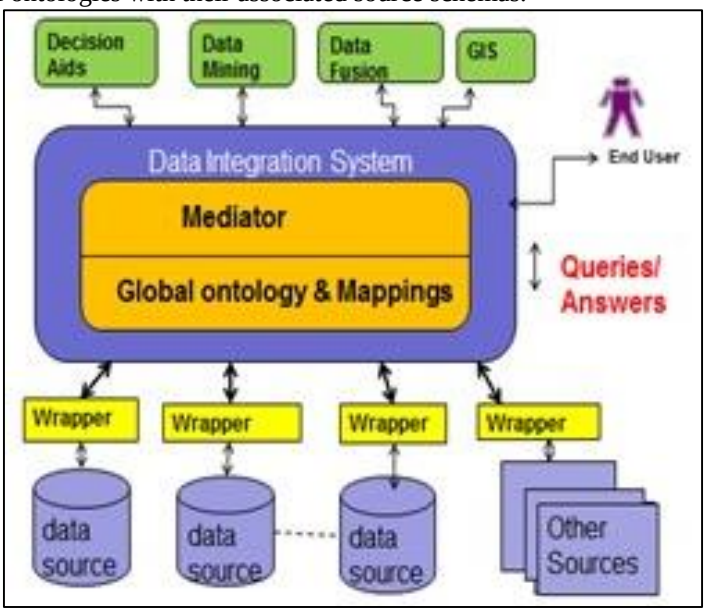

Fig. 4: Local to global Ontology mapping [4]

This method tries to solve conflicts at both the level at data as well as schema level. Data level conflicts like data value having different meaning, data representation have different format, mayuse different data unit. Schema level conflict include like name, generalization, aggregation etc.

1) Advantages:

- Interoperability, data accurateness and consistency are achieved.
- Tools used to automate partly the data integration task and reduce efforts
- Reduce the amount of integration decisions and the number of rules

2) Disadvantages:

- Ontology reuse is cost-effective instead of building a new one from scratch
- Integrating vast amount of information from different sources is a difficult, complex and demanding task
- In the absence of a unique concept identifier, the integration process will remain semiautomatic driven by a domain expert

E. Ontology Based Data Warehouse Schema Design:

This method uses hybrid methodology [13][14] to derive the multidimensional model. The requirements and the data source are represented using ontology. Using ontology reasoning capability the multidimensional elements such as fact, measure and dimension are derived from the data source.

The concepts from the requirements are matched with the multidimensional elements to filter the results. After including user suggestion the resultant conceptual model is represented graphically. The tool also facilitates to convert the conceptual to the logical model. Finally quality of the logical multidimensional model is assessed. The OBDWSD (ontology based data warehouse schema design) tool is developed using Java (j2sdk 1.4.2) and Jena 2.1, the Java API for ontology development and processing. We make use of ontology matching algorithm to perform matching between requirements and data source ontology

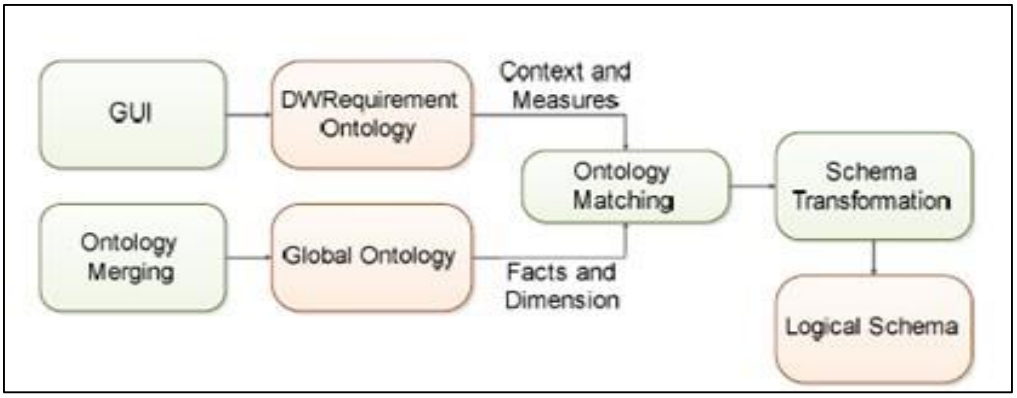

Fig. 5: Local to global Ontology mapping

This tool has GUI features which help the designer to assist in generating the multidimensional logical schema. The different components that are integrated in this tool is shown in Figure. 5.

1) Advantages:

- DWRequirement mapped to Ontology
- Automatically deriving multidimensional elements present in the data source ontology.
- Formally match requirements with the data source to filter results.
- Generation of Logical Schema.

III. COMPARATIVE STUDY

Data warehouse schema design in the earlier was carried out mainly considering relational sources. At present, due to increasing amount of web source and the popularity of semantic web lead to different ontological approaches for data warehouse schema design. These approaches followed supply driven or demand driven and only a few follow the hybrid approach. As shown in the table 1 ontology based data warehouse schema design tool is more suitable than other approaches. OBDWSD is a fully automatic method that is there is no need of human intervention while designing the ontology and integrating sources. In this approach while mapping an ontology user requirement and data warehouse requirement mapped into global ontology. Designing of the system based on hybrid approach, where both the supply driven and demand driven both the methods are considered while collecting requirements from the user. The architecture supported by this is agent based architecture. This method is fully automatic method hence there is no need for human intervention to design ontology.

IV. CONCLUSION

In this work, we have presented both an analysis and a comparison of four systems that use ontologies to solve the problems involved in data integration. In order to do so, we have used a conceptual framework with three main categories: architecture, semantics heterogeneity and query resolution. Based upon our comparison, we have found some elements in common and also original aspects of the systems.

Features\Methods	DOMM	HDSDQ	LTOG	OBDWSD
Automation	Semi-automatic	Semi-automatic	Semi-automatic	Fully
Mapping Method	Data Bases To Ontology Classes	ER To Metadata To Ontology	Data/Schema Mapped To Ontology Then Ontology To Ontology	User Requirement And Data Warehouse Requirement Into Global Ontology
Design Approach	Hybrid	Hybrid	Demand-driven	Hybrid
Architecture Type	Agent Based	Wrappers	Wrapper, Mediator	Agent Based
Ontology Use	Global Ontology	Multiple Ontology	Global Ontology, Hybrid Ontology	Global Ontology
Language	Classic/OWL	Classic / LOOM / OWL	Loom /OntoDSL	OWL
Tool	Etl,d2o	Dc,mafra,rdft	Protégé	Etl, jena 2.1, Protégé, Pellet Reasoner API
Formal Algorithm	No	No	No	Yes
Automation	Data Bases To Ontology Classes	ER To Metadata To Ontology	Data/Schema Mapped To Onto Then Onto To Onto	User Requirement And Data Warehouse Requirement Into Global Ontology

Table 1: Comparison of Different Techniques

As this brief survey shows, many issues that ontology researchers in semantic integration deal with are very similar to the issues that database and information-integration researchers have been addressing. Some of the approaches are also similar although the ontology community relies more heavily on the higher expressive power of ontology languages and on reasoning techniques. With ontologies, using a common upper ontology or reference ontology to improve the integration problem is also a common approach.

The two communities can certainly share and reuse the techniques that they have developed in their respective domains. In fact, there has been a certain convergence trend where schema-matching approaches for example employ more expressive components of schema definitions in their techniques.

REFERENCES

[1] Michael J. Herold, SatyajeetRaje, Jay Ramanathan, Rajiv Ramnath, and ZheXu " Human Computation Recommender for Inter-Enterprise Data Sharing and ETL Processes " 2012 Department of Computer Science & Engineering The Ohio State University Columbus, Ohio 43210–1277

[2] Jinpeng Wang, Jianjiang Lu, Yafei Zhang, Zhuang Miao and Bo Zhou, "Research on Heterogeneous Data Sharing Based on Ontology " 2012

[3] Shubhra S. Ray et al. (2009). "Combining Multi-Source Information through Functional Annotation based Weighting: Gene Function Prediction in Yeast". IEEE Transactions on Biomedical Engineering 56 (2): 229–236. doi:10.1109/TBME.2008.2005955. PMID 19272921.

[4] Lihong Jiang1, Junliang Xu1, Boyi Xu2, "An Automatic Method of Data Warehouses Multidimensional Modeling for Distributed Information Systems" Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design 978-1-4577-0387-4/11 2011 IEEE

[5] Lihong Jiang, HongmingCai Shanghai, China "A domain ontology approach in the ETL process of Data Warehousing" Nov. 2010 pp 30-35

[6] AgustinaBuccella*, Alejandra Cechich* and Nieves R. Brisaboa† "Ontology-Based Data Integration Methods: A Framework for Comparison", 2010 pp165-168

[7] LeylaZhuhadar and OlfaNasraoui, Elizabeth Romero, " Multi-model Ontology-based Hybrid Recommender System in E-learning Domain", 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops 20

[8] Shaker H. Ali El-Sappagh a,*, Abdeltawab M. Ahmed Hendawi b, Ali Hamed El Bastawissy "A proposed model for data warehouse ETL processes", 2011

[9] S. Naciri, M. Pouly, J. Binggeli, and R. Glardon, "Using the Generic Product Model for storing and sharing ERP data," CSCWD'09, pp.618-623

[10] Sheth.Changing focus on interoperability in information system: Form system, syntax, structure to semantics [C].Interoperating Geographic Information Systems, Boston: Kluwer Academic Publisher, 1999, PP: 13-14

[11] YannisKalfoglou, and Marco Schorlemmer, Ontology mapping: the state of the art, The Knowledge Engineering Review Journal, Vol 18, No. 1, pp.1-31

[12] Shaker H. Ali El-Sappagh a,*, Abdeltawab M. Ahmed Hendawi b, Ali Hamed El Bastawissy "A proposed model for data warehouse ETL processes" ,2011 ,pp 91–104

[13] M.Thenmozhi1, K.Vivekanandan2 "A Tool for Data Ware house Multidimensional Schema Design using Ontology", 2013 pp. 33-39