ONTOLOGY BASED DATA INTEGRATION WITH USER FEEDBACK

Devini.K, M.S. Hema

Abstract-Many applications need to access multiple heterogeneous data sources. The integration of these data sources raises several semantic heterogeneity problems. In existing systems, it is difficult to provide the high quality result to the end users due to the heterogeneity, inaccuracy of the facts in data sources. This paper is proposed to resolve semantic heterogeneity problem in integration by using ontology and to provide the quality results to the end user and improve the quality of data sources of data integration system by using user feedback. Online e-shopping application is taken for experimentation. The experimentation results shows that the response time and precision of the data is improved.

Index Terms- Data integration, Ontology, User Feedback, cluster, cache, heterogeneity.

1. INTRODUCTION

In many domains, the number of data providers and amount of available data is increasing tremendously. However, users usually require an integrated view of the data available from heterogeneous data sources. Therefore, integration issues are attracting ever more attention. Data integration refers to combining data in such a way that a homogeneous and uniform view is presented to users. In the case of Data Integration, the problem is particularly complex due to the integration of data coming from multiple sources and possibly having different quality.

The three types of data integration methods are 1. Data consolidation 2.Data propagation and 3.Data federation[1]. The data federation provides a single virtual view for two or more data sources. When a business application issues a query against this virtual view, a data federation engine retrieves data from the appropriate data sources and integrates it. By definition, data federation always pulls data from source systems on an on-demand basis. Enterprise information integration is an example of a technology that supports a federated approach for data integration. The main advantage of data federation is Data are not moved or copied from source systems, so additional storage is not required.

The creation of virtual view among different data sources is a challenging task due to presence of various types of heterogeneities. Among those heterogeneities, the semantic heterogeneity is difficult to handle. Semantic heterogeneity is due to the difference in the interpretation of the meaning.

- *Devini k is currently pursuing masters degree program in computer science and engineering in Anna University, India ,E-mail: rdevinikrishnamoorthy@gmail.com*
- *M.S.Hema is currently working Associate professor in computer science and engineering in Kumaraguru College of Technology, India, E-mail: hema.m.s.cse@kct.ac.in*

This semantic heterogeneity can be resolved by using ontology.

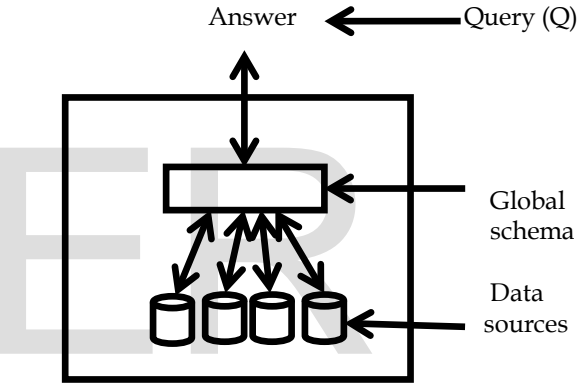

Figure1 Wrapper Architecture

A data integration system (DIS) is composed of three elements. They are global schema, set of source schemas (including schemas of all sources), mapping between the global schema and the source schemas. Figure1 shows that Wrappers have the main task of mapping local schema with global schema and mapping query from global to local schema[2]. There are three basic approaches to create mapping between local schema and global schema. The first approach, called global-as-view (GAV), requires the global schema to be expressed as views over the local schema of data sources. The second approach, called local-as-view (LAV), requires each data source to be expressed as views of queries over the global schema. A third approach is called global-local-as-view (GLAV), and it is a mixture of the two; it combines the GAV and LAV approaches in such a way that queries over the sources are put into correspondence with queries over the global schema.

The Quality of data (QoD) is a multidimensional, complex concept. Some of the significant quality dimensions are Data completeness, uniqueness, consistency, freshness and accuracy. These dimensions are important for quality of data. Data Completeness concerns the degree to which all data relevant to an application domain has been recorded in an information system. It expresses that every fact of the real world is represented in the information system. Coverage and density are the two aspects of completeness. Coverage describes whether all required entities for an entity class are included.Density describes whether all data values are present (not null) for required attributes.

Data Uniqueness states that two or more values do not conflict each other. Data Consistency expresses the degree to which a set of data satisfies a set of integrity constraints. Data is said consistent if it satisfies these constraints .The most common constraints checks for null values, key uniqueness and functional dependencies.

Data Accuracy is concerned with the correctness and precision with which real world data of interest to an application domain is represented in an information system. In existing systems [7], it is difficult to provide the quality result for the end users as because of the multiple data sources. The quality of the data conveyed to users is an important problem, which is closely related to the success of data integration system. In this paper, the quality of the data integration result is improved by getting the feedback from the user.

In information retrieval [8] the relevance feedback is a form of user feedback that is well studied, but its aim is not to improve data, it contextualizing the queries. The goal of user feedback is to improve quality by updating the database according to the feedback. Linguistic feedbacks are obtained from the user, if the feedback is crossing the threshold values, then the feedback are clustered to realize the user requirements.If the feedback is low, medium then the reason for the corresponding feedbacks are also got from the user to improve the quality of data sources and the quality of result. The feedback, query, result and reason for the low, medium feedback are stored in cache.

2. RELATED WORK

Many Researches showing the importance of handling semantic heterogeneity problem, and many techniques have been proposed to improve the heterogeneity problems.

Yi Peng, Yong Zhang[11] proposed One of the fundamental problems in incident information management is how to integrate and analyze heterogeneous incident data and provide intelligent decision support to decision makers(DMs). This study proposes a conceptual framework for incident information management to support information integration, intelligent data analysis, and multi-criteria decision making. It develops a three-level framework for incident information management, including heterogeneous data integration, data mining and multi-criteria-decision-making (MCDM), and collaboration tools. Data integration level provides a distributed heterogeneous data interface that integrates various data sources and a unified data interface that facilitates differentiated services to upper application modules.

Maurice van Keulen[12]proposed probabilistic integration approach aims at reducing the development effort needed for such applications by allowing some semantic uncertainty to remain in the data, while still being able to meaningfully use this data. The developer is only required to provide a few knowledge rules and rough estimations for thresholds to produce a usable initial integration. The main contribution of this paper is a thorough experimental investigation of the effects and sensitivity of rule definition, threshold tuning, and user feedback on the integration quality. But if incorrect feedback is given, correct information may get lost.

Khalid, Norman, Alvaro[13] this paper that treating feedback as a first class citizen presents several advantages. It presented a straightforward model for describing different kind of feedback that has been considered in the information integration literature [14]. Furthermore, it identified issues that underlie feedback management in information integration systems.

In InfoSleuth[15], a set of agents collaborate with each other for information discovery and retrieval in a dynamic environment. Ontologies built in InfoSleuthare used to represent semantic concepts consistent across all the InfoSleuth agents, and a common vocabulary or domain model facilitates agent communication.

3. PROPOSED WORK

This paper proposes Ontology Based Data Integration with User Feedback Architecture for data integration as shown in Figure 2. This architecture has four layers i)Database Layer, ii)Federation Layer, iii) Caching and Feedback Layer,iv)User Interface Layer.

3.1 Database Layer

Database Layer Contains all the data sources which are participating in data integration. The data sources return the result to the given query.

3.2 Federation Layer

Federation layer contains two layers Ontology Layer and Query Decomposition Layer.

The ontology layer contains the two sub layers i) the Local Ontology Creates from local schema of the data sources. A web ontology language (OWL) is used to describe the ontology. It defines OWL classes, relationship, properties and constraints. Each concept and attribute in the local ontology will be mapped to the global ontology by using mapping rule. ii) Global ontology is constructed using LocalasView[LAV] approach. The LAV combines the concepts and attributes with same semantics and creates global ontology using Protégé tool 4.2. II) In Query processing layer the Query processing has the following two services, i) Query Decomposition service, ii)Query Execution service.

Query Decomposition Service In Decomposition the global query Q is decomposed into Q1,Q2,…,Qn and passed to respective data sources. ii) The Query Execution Service returns the result from multiple data sources, the data sources returns a set of results R1,R2,….,Rn. Each result may have multiple records. It integrates data from multiple data sources and to provide a unique result to the user.

3.3 Caching and Feedback Layer

Caching service caches the Query and result. Clustering service group the feedback for the output data.

The Caching and Feedback Layer contains i) Query and Result Caching Service, query and result stored in cache, ii) In Result update service the query posed by the user is verified in the cache and the result is displayed directly from the cache, if the posed query is already in the cache. Otherwise the query is processed and stored along with the result in the cache, iii) In Feedback service if the user expecting to improve results and data quality, the user gives the Feedback for the result provided as low, medium, high. If the user rates the results medium or poor the reasons for such rating are also obtained. So that the user requirements are better realized and the system is improved to provide quality results, iii) In Clustering Service the similar feedback suggestion grouped for a single query.

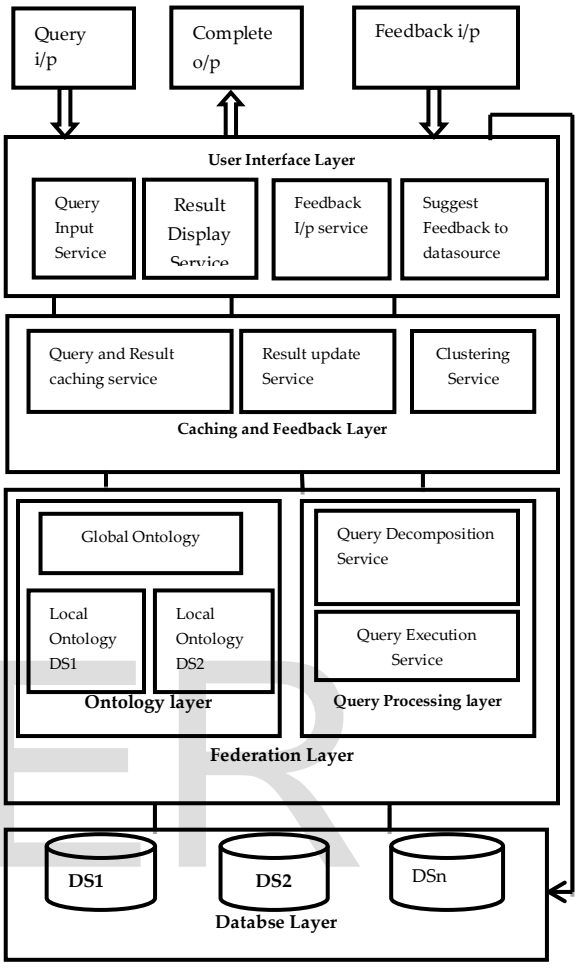

Fig 2 Architecture of Ontology Based Data Integration with user feedback

In data integration User Feedback is a growing and our approach uses linguistic measure i) Low(L), ii) Medium (M), iii) High (H) to capture the user expectations to improve the quality of the data. When the results presented to the user, it satisfies the user needs then theuser feedback value is high. If the result doesn't satisfy the user needs then the user feedback value is low. If it is not satisfied but somewhat relate to the user needs then the value is medium. The queries and the respective attributes present in the user feedback and result are stored in as shown in Table 1.

Table 1 User feedback database

	L	M	H	RSL	RSM
Q_1			5		
Q_2	3			conflicting result, incorrect result, duplicate result	
Q_3		4			duplicate result, incorrect answer, low response time,
….					
Q_n					

Q1, Q2, Q3, …Qn- Query, L-low, M-medium, H-high, RSL-Reason for low feedback, RSM-Reason for medium feedback, R-Result.

This Database holds initially '0' for all attributes not present in the result, feedback, reason for particular query. If one user gives the query (Q1) result is high, the value of the attribute will be incremented by 1. Another user also gives the same feedback for the same query (Q1) the value will be incremented by 1 (i.e) now high attribute value is 2.

3.4 User Interface Layer

The user gives query and gets the corresponding result through the user interface layer. This layer consist i) Query input module accepts query(SPARQL query) from user and it ported for processing, ii) Result Display Module displays the integrated quality result to user for the corresponding posed query, iii) Feedback Input module get the linguistic feedback from the user like low (L), medium (M), high (H) for the corresponding output. If the user feedback is low or medium, get the reason for the feedback from the user, iv) suggestion feedback module raise suggestion to the data source if it is above the threshold value.

4. IMPLEMENTATION

For experimentation the following services are implemented local ontology service, global ontology service, result integration service and result cache service. This implementation is done on shopping data set for online e-shopping application.

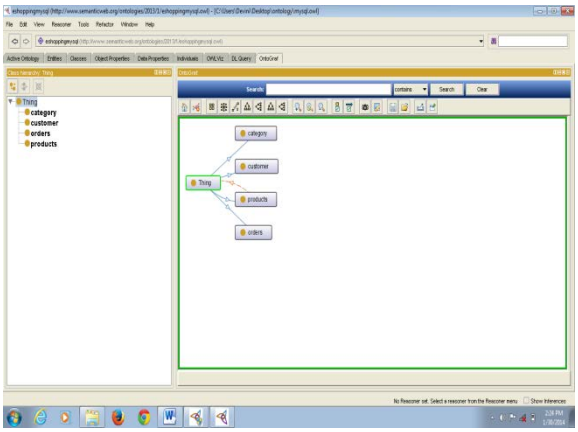

Figure 3 Ontograph of table in data source

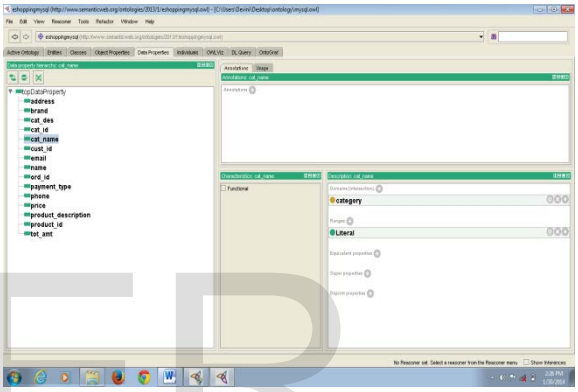

Figure 4 Classes, properties and relationship

5.CONCLUSION

In this paper implemented ontology based data integration with data quality by using user feedback. In ODBI-U (Ontology Based Data Integration with User feedback) architecture semantic heterogeneity problems are resolved. The Quality of the query result improved by using user feedback and also this architecture suggested to the data source for improve data source quality. The result of this proposed work improved the response time by using cache and improved the precision of the result in the data integration system by using user feedback. Improvisation of data source quality depending on user feedback and enhancement of caching service with respect to storage cost can be carried out as future work.

REFERENCES

[1] Olga Brazhnik, John F. Jones "Anotomy of data integration", ACM Nov 2007.

[2]VipulKashyap, AmitSheth," Semantic and schematic similarities between database objects: a context-based approach",Springer,1996.

[3]M. Abdul Rehman,StefanJablonski, Bernhard Volz,"An Ontology Based Approach to Automating Data Integration in Scientific Workflows",ACM,Dec 2009.

[4]AmitSheth and James A. Larson."Federated database systems for managing distributed, heterogeneous and autonomous databases". ACM Computing Survey, 22(3):183 236, 1990.

[5] ZohraBellahsene, SalimaBenbernou, H«el „eneJaudoin" FORUM: A Flexible Data Integration System Based on Data Semantics"∗SIGMOD 2010 (Vol. 39, No. 2).

[6]Jainin Wang, "A Quality Framework Data Integration", In Interest Group on Management of Data, pp- 338, 2010.

[7]Bradji, M.Boufaida, "User Expectation Feedback Consistency As a First Step For a Better Data Quality", In JATIT, Volume-1,pp-68, 2011.

[8]Monica Scannapieco, AntoninoVirgillito, Carlo Marchetti, Massimo Mecella, and Roberto Baldoni. "The DaQuinCIS Architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems. Information Systems", University of DegliStudi Di Roma, 29(7):551–582, 2004.

[9]AnHai Doan Robert McCann, "Building Data Integration Systems: A Mass Collaboration Approach"Computing survey,2010.

[10] Shawn R. Jeffery, Michael J. Franklin, Alon Y. Halevy "Pay-as-you-go User Feedback for DataspaceSystems",SIGMOD june,2008.

[11] Yi Peng, Yong Zhang, Yu Tang, Shiming Li" An incident information management framework based on data integration, data mining, and multi-criteria decision making"ELSEVIER, Nov 2010.

[12]Maurice van Keulen, Ander de Keijzer "Qualitative Effects of Knowledge Rules and User Feedback in Probabilistic Data Integration",ACM 11, june,2009.

[13] Khalid Belhajjame, Norman W. Paton, Alvaro A. A. Fernandes, Suzanne M. Embury, "User Feedback as a First Class Citizen in Information Integration Systems", ACM, jan 2012.

[14]Maurizio Lenzerini,"Data integration: A theoretical perspective", ELSEIVER, 2009.

[15]Nodine, M., Ngu, A.H.H., Cassandra, A., and Bohrer, W.G Scalable Semantic Brokering over Dynamic Heterogeneous Data Sources in InfoSleuth. *IEEE TRANSACTIONS ONKNOWLEDGE AND DATA ENGINEERING*, 15, 5 (2003),1082-109.