JCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-2, Issue-11 E-ISSN: 2347-2693

A Survey: Different Approaches to Integrate Data Using Ontology and Methodologies to Improve the Quality of Data

Sowmya Devi L[1*], Jai Barathi B[2], Hema M.S.[3] and S. Chandramathi [4]

[1*,2,3,4]*Computer Science & Engineering, Kumaraguru College of Technology, India*

www.ijcaonline.org

Received: Oct/26/2014 Revised: Nov/09/2014 Accepted: Nov/20/2014 Published: Nov/30/2014

***Abstract*—** This In today's world, the amount of data is increasing tremendously. In order to analyze data and make decisions, data residing at different sources are integrated. Data integration is an approach to integrate data from different data sources. Data federation is a data integration strategy used to create integrated virtual view. This paper deals with various approaches of data integration to resolve semantic heterogeneity using ontology. Various ontology based data integration techniques are reviewed and issues are summarized. Different metrics and approaches are also discussed to improve the quality of the data.

***Keywords*—**Data Integration, Ontology, Semantic heterogeneity, Data quality

I. INTRODUCTION

Data integration is an approach to integrate data from multiple data sources. The three different approaches for data integration are data consolidation, data propagation and data federation. Data federation is an approach which creates a virtual view of the resultant database. Ontology based data integration is an effective approach and gives the better results. Ontology is the formal explicit specification of the shared conceptualization. The different approaches for integrating the data and methodologies to improve the quality of data are discussed in this survey paper.

II. DATA INTEGRATION

Data integration is carried out to integrate data from various heterogeneous data sources. Integration of informational data is carried out by various integration approaches [1]. They various integration approaches are as follows

A. Manual Integration

Here, users directly interact with all relevant information systems and manually integrate selected data. That is, users have to deal with different used interfaces and query languages.

B. Common user interface

In this case, the user is supplied with a common user interface (e.g. a web browser) that provides a uniform look and feel. Data from relevant information systems is still separately presented so that homogenization and integration of data yet has to be done by the users.

C. Integration by application

This approach uses integration applications that access various data sources and return integrated results to the user. This solution is practical for a small number of component systems

D. Integration by middleware

Middleware provides reusable functionality that is generally used to solve dedicated aspects of the integration problem, example as done by the SQL middleware

E. Uniform data access

In this case, a logical integration of data is accomplished at the data access level. Global applications are provided with a unified global view of physically distributed data, though only virtual data is available on this level. This global view of physically integrated data can be time consuming since data access, homogenization and integration have to be done at run time.

F. Common data storage

Here, physical data integration is performed by transferring data to new data storage; local sources can either be retired or remain operational.

During the integration of informational data from various data sources, resolving heterogeneities remains as a challenging task. The heterogeneities available in the data bases make data integration a tougher task. The various heterogeneities are as follows.

- Structural heterogeneity: It involves different data models.
- Systematic heterogeneity: It involves hardware and operating system.
- Syntactical heterogeneity: It involves different languages and data representations
- Semantic heterogeneity: It involves different concepts and interpretations. Semantic heterogeneity deals with three types of concepts.

- Semantically equivalent concepts
- Semantically related concepts
- Semantically unrelated concepts

III. ONTOLOGY BASED DATA INTEGRATION

There are several methods created to address the problem of dealing with different concepts and interpretations. Use of ontology is used as one of the methods to resolve heterogeneities. Ontology is defined as "the formal explicit specification of the shared conceptualization". Data integration carried out using the ontology is of three types. They are single ontology, multiple ontology and hybrid ontology [2]. A comparative study is made among various ontology approaches which are shown in the Table 1.

Table 1: Comparison of ontology approaches

Approaches	Implementation effort	Semantic heterogeneity	Adding/ removing of sources
Single ontology	Straight forward	Similar view of a domain	Need for some adaption in the global ontology
Multiple ontology	Costly	Supports heterogeneous views	Providing a new source ontology; relating to other
Hybrid ontology	Reasonable	Supports heterogeneous views	Providing a new source ontology

IV. DATA MODELLING

The data integration systems are characterized by an architecture based on global schema and a set of sources. The sources contain the real data while the global schema provides a reconciled, integrated and virtual view of the underlying sources. There are two basic approaches proposed [3].

A. *Global-as-view (GAV) Model*

In this model, the global schema is defined by having one or more views over the source schemas for each class. In this approach, changes in information sources or adding a new information source requires mapping between the global and source schemas.

B. *Local-as-view(LAV)Model*

In this model, the source database is modeled as a set of views over an underlying global schema. The advantage of this model is that new sources can be added easily when compared to GAV. However the query rewriting process is complex because the system has to choose from a set of choices to determine the best possible rewrite.

V. DIFFERENT SYSTEMS

In this survey paper, we have presented the analysis and comparisons of seven systems that use ontology to solve the problems involved in data integration. In order to do so, a conceptual framework with three main categories has been created. They are architecture, semantic heterogeneity and query resolution. The seven systems are as follows.

A. *SIMS [Search in Multiple Sources]*

In [4], authors Arens. Y, Hsu. C, Knoblock. C has discussed the architecture, semantic heterogeneity and query resolution of the SIMS system. SIMS was created assuming dynamic information sources, i.e. changing information sources, availability of new information, etc. the sources can be databases and information sources such as HTML pages. The architecture is based on the wrapper/mediator. A wrapper is used to translate a data set description into a query, which is submitted to the source. Mediator is used to retrieve and process data. A global ontology approach is used in the SIMS. The ontology is represented in the Loom language.

Users make a query in terms of the global ontology without knowing the terms or languages used by the underlying information sources. Queries are written in high level languages. The first step to answer a query is transforming it into another query expressed in terms of concepts that correspond to information sources. The four reformulation operations are as follows.

- Select-Information-Source
- Generalize- Concept
- Specialize concept
- Decompose relation

SIMS uses the Semantic Query Optimization that can speed up database query answering by using knowledge intensive reformulation.

B. *OBSERVER [Ontology Based System Enhanced with Relationship for Vocabulary hEterogeneity Resolution]*

In [5], OBSERVER is an approach that proposes managing multiple information sources through ontologies. OBSERVER uses the concept of data repository, which might be seen as a set of entity types and attributes. The architecture is based on wrappers, ontology servers and an IRM (Inter- ontology Relationship Manager). OBSERVER is classified as multiple ontology approach. In this system, each information source is represented by one ontology, thus a modification or addition of information to some source will only impact on the related ontology and on the IRM. Users use any language based on description logics such as CLASSIC or Loom.

The query construction is carried out by the user. It is followed by the access to underlying data and the controlled query expansion to new ontologies steps.

C. DOME [Domain Ontology Management Environment]

In [6], DOME is focused on ontology development by using software reverse engineering techniques. The most important architectural components are wrappers, a set of tools for extracting and defining ontologies and mappings between them, the mapping server and the ontology server. The DOME system uses the multiple ontology approach. DOME uses XRA as a tool to generate ontologies.

D. KRAFT [Knowledge Reuse And Fusion/Transformation]

In [7], KRAFT was conceived to support configuration design of applications among multiple organizations with heterogeneous knowledge and data models. It uses the concept of "Knowledge fusion" to denote the combination of knowledge from different sources in a dynamic way.

E. COIN [Context Interchange]

In [9], COIN system is with a goal of achieving semantics interoperability among heterogeneous information sources. It has mediator based architecture. This COIN technology uses a hybrid ontology approach.

F. GARLIC

It addresses large scale multimedia information systems by considering specialized component systems to store and search for particular data types like image management systems. Garlic provides an object-oriented schema to applications, interprets object queries, creates execution plans for sending pieces of queries to appropriate data servers, and assembles query results for delivery back to the applications.

Even changes of capabilities do not affect the mediator. Garlic requires quite powerful wrappers, since query execution depends on a interactive communication between mediator and wrappers about the component's capabilities. Table 2 gives the comparison of different systems used in ontology based data integration.

Systems	**Information sources**	**Architecture type**	**Ontology use**	**Languages**	**Query**
SIMS	HTML pages	Wrapper/mediator	Single ontology	Loom	Query subsumption
OBSERVER	HTML Pages, databases and files.	Wrappers, ontology servers and IRM	Multiple ontology approach	CLASSIC or Loom	Cost based query optimization
DOME	Structured and semi- structured data sources	Wrappers, mapping server and ontology server	Multiple ontology approach	CLASSIC	Cost based query optimization
KRAFT	Knowledge bases	Wrappers, mediators, facilitators and user agents	Hybrid ontology approach	Classical frame based representational language	Constraint based query
COIN	traditional databases and semi structured sources	Mediator based architecture	Hybrid ontology approach	F- Logics	Cost based query optimization

Table 2: Comparison of different systems for ontology based data integration

VI. METHODOLOGIES FOR DATA QUALITY ASSESSMENT

The goal of this survey paper is to provide a systematic and comparative description of different methodologies of data quality assessment. The classifications of quality dimensions are provided. By analyzing these classifications it is possible to define a basic set of quality dimensions, including accuracy, completeness, consistency and timeliness. The different methodologies are as follows.

A. The TDQM (Total Data Quality Management) Methodology

In [10], the TQDM methodology was the first general methodology published in the data quality literature [Wang 1998]. The objective of TDQM is to extend to data quality, the principle of Total Quality Management (TQM). TDQM proposes a language for the description of information production (IP) processes, called IP-MAP. IP-MAP has been variously extended, towards UML and also to support organizational design.

TDQM's goal is to support the entire end-to-end quality improvement process, from requirement analysis to implementation. TDQM cycle consists of four phases that implement a continuous quality improvement process: definition, measurement, analysis and improvement.

B. The DWQ (Data Warehouse Quality) Methodology

In DWQ heterogeneous information sources are first made accessible in a uniform way through extraction mechanisms called wrappers, and then mediators take on the task of information integration and conflict resolution [11]. The resulting standardized and integrated data is stored as materialized views in the data warehouse.

C. TIQM (Total Information Quality Management) Approach

In [12], TIQM methodology has been proposed to support data warehouse projects. The methodology assumes the consolidation of operational data sources into a unique integrated database, used in all types of aggregations performed to build the data warehouse. The goal is to improve the data quality level.

D. AIMQ (A Methodology for Information Quality Assessment)

In [13], the AIQM methodology is the only information quality methodology focusing on benchmarking, that is an objective and domain independent technique for quality evaluation. Gap Analysis Technique is advocated as a standard approach to conduct benchmarking and interpret results.

E. CIHI (Canadian Institute for Health Information)

In [14], the CIHI has implemented a method to evaluate and improve the quality of Canadian Institute for Health Information data. In the CIHI, the main issue is the size of databases and their heterogeneity. It also proposes a large set of quality criteria to evaluate heterogeneity. CIHI Data Quality strategy proposes a two phase approach. The first phase is definition of a Data Quality Framework and the second is in depth analysis of the most frequently accessed data.

F. The DQA (Data Quality Assessment) Methodology

In [15], the DQA methodology has been designed to provide the general principles guiding the definition of data quality metrics. The objective metrics are classified into task dependent and task independent.

G. The IQM (Information Quality Measurement)

In [16], the fundamental objective of the IQM methodology is to provide an information quality framework tailored to Web data. In particular, IQM helps the quality based selection and personalization of the tools that support webmasters in creating, managing and maintaining websites.

H. The ISTAT (The Italian National Bureau of Census) methodology

ISTAT suggests how to resolve heterogeneities among data managed by different public agencies by adopting a common model for representing the format of exchanged data, based on the XML markup language [17]. In this way, the comprehension of heterogeneities among agencies is made easier, while the solution of such heterogeneities is left to bilateral or multilateral agreements.

I. AMEQ (Activity-based Measuring and Evaluating of Product information Quality) methodology

The main goal of AMEQ methodology is to provide a rigorous basis for Product Information Quality (PIQ) assessment and improvement in compliance with organizational goals [18].

J. COLDQ (Cost-effect Of Low Data Quality)

The fundamental objective of COLDQ methodology is to provide a data quality scorecard supporting the evaluation of the cost effect of low quality data [19].

K. DaQuinCIS (Data Quality in Cooperative Information System) Methodology.

In DaQuinCIS, instance-level heterogeneities among different data sources are dealt with by the DQ broker. Different copies of the same data received as responses to the request are reconciled by the DQ broker, and a best-quality value is selected [20].

L. QAFD (Quality Assessment of Financial Data) Methodology.

In [21], the QAFD methodology has been designed to define standard quality measures for financial operational data and thus minimize the costs of quality measurement tools. The QAFD selects the most relevant financial variables. Selection is based on knowledge from previous assessments, according to their practical effectiveness. Then the most relevant data quality dimensions are identified in this phase and data quality rules are produced.

M. CDQ [Complete Data Quality]

In [22], CDQ follows an approach similar to ISTAT with more emphasis on the autonomy of organizations in the cooperative system. In fact, the resolution of heterogeneities proposed as best practices are performed through record linkage on a very thin layer of data, namely the identifiers.

Methodologies	Data quality dimension	Type of data	Extensible to other dimensions and metrics
TDQM	Timeliness, security	Monolithic, distributed	Fixed
DWQ	Correctness, traceability	Strongly focused on data warehouse	Open
TIQM	Concurrency of redundant data	Focused on monolithic and distributed	Fixed
AIMQ	Freedom from errors	Monolithic	Fixed
CIHI	Linkage ability	Monolithic, distributed	Fixed
DQA	Ease of manipulation	Distributed is implicitly considered	Open
IQM	Accuracy, interactivity	Strongly focused on web	Open
ISTAT	Accuracy, completeness	Monolithic, distributed	Fixed
AMEQ	Unambiguity, consistency	Monolithic	Open
COLDQ	Accuracy, completeness	Monolithic	Fixed
DaQuinCIS	Consistency, currency	Monolithic, distributed	Open
QAFD	Syntactic/ semantic accuracy	Monolithic	Fixed
CDQ	Syntactic/ semantic accuracy	Monolithic, distributed	Open

Table 3: Comparison of Different Data Quality Methodologies

VII. CONCLUSION

Resolving semantic heterogeneity is the challenging task in data integration. In this paper we have discussed several systems that use ontology to solve the problem involved in data integration. Different methodologies are also used to improve the quality of the integrated data.

REFERENCES:

[1] Batini C, Lenzerini M, Navathe S B, "A comprative analysis of methodologies for schema integration",Journal of ACM Computing Surveys, (CSUR 1986),Volume-**18,** Issue-**4**, page no (**2-6**), January **1986**.

[2] Yushui Geng,Xiangcui Kong, "The Key Technologies of Heterogeneous Data Integration System Based on Ontology", International Workshop on Education Technology and Training, page no (**723-725**), January **2008.**

[3] Maurizio Lenzerini, "Data Integration: A Theoretical Perspective", Proceedings of PODS, page no (**233-246),** December **2002**.

[4] Arens, Y., Hsu, C., Knoblock, C.A, "Query processing in the SIMS Information Mediator". Advanced Planning Technology, Austin Tate (Ed.), AAAI Press, Menlo Park, CA, page no (**61-69**), **1996**.

[5] Mena, E., Kashyap, V., Sheth, A. and Illarramendi, A. "Observer: An approach for query processing in global information systems based on interoperation across pre-existing ontologies", Kluwer AcademicPublishers,Boston.http://citeseer.nj.nec.com/mena96observer.html, 1-49, **2000**.

[6] Cui, Z. and O'Brien, P. "Domain Ontology Management Environment". In Proceedings of the

33rd Hawaii International Conference on System Sciences, **2000**.

[7] Gray, P.M.D, Preece, A., Fiddian, N.J. and colab, "KRAFT: Knowledge Fusion from Distributed Databases and Knowledge Bases", Proceedings of 8th International Workshop on database and Expert Systems Applications (DEXA'97), **1997**.

[8] Woelk, D., P. Cannata, M. Huhns, W. Shen, and C. Tomlinson. Using Carnot for Enterprise Information Integration. Second International Conf. Parallel and Distributed Information Systems, page no (**133-136**), January **1993.**

[9] Goh, C.H., Bressan, S., Siegel, M. and Madnick, S. E. "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information". ACM Transactions on Information Systems, Vol. **17(3)**, pp (**270–293**), **1999.**

[10] Wand, R. A product perspective on total data quality management. Comm. ACM 41, 2, **1998**.

[11] Jeusfeld, M.,Quix, C., and Jarke,M. "Design and analysis of quality information for datawarehouses". In Proceedings of the 17th International Conference on Conceptual Modeling, **1998**.

[12] English, L. "Improving Data Warehouse and Business Information Quality". Wiley & Sons, **1999**.

[13] Lee, Y.W., Strong, D. M., Kahn, B. K., Andwang, R. Y. "AIMQ: A methodology for information quality assessment". Inform. Manage. 40, 2, pp (**133–460**), **2002.**

[14] Long, J. and Seko, C, "A cyclic-hierarchical method for database data-quality evaluation and improvement. In Advances in Management Information Systems-Information Quality Monograph (AMISIQ)", April **2005**.

[15] Monograph, R. Wang, E. Pierce, S. Madnick, and Fisher C.W. Pipino, L., Lee, Y., and Wang, R., "Data quality assessment". Commun. ACM 45, 4, **2002.**

[16] Eppler, M. and Munzenmaier, P, Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology. In Proceedings of the 7th International Conference on Information Systems (ICIQ).ISTAT, **2002**.

[17] Guidelines for the data quality improvement of localization data in public administration (in Italian). www.istat.it.**2004**.

[18] Su, Y. and Jin, Z. 2004. "A methodology for information quality assessment in the designing and manufacturing processes of mechanical products", In Proceedings of the 9th International Conference on Information Quality (ICIQ). Page no (**447–465**), December **2004**.

[19] Loshin, D. "Enterprise Knowledge Management - The Data Quality Approach. Series in Data Management Systems", Morgan Kaufmann, chapter **4**, **2004**.

[20] Scannapieco, M., M.Virgillito, Marchetti, M., Mecella, M., and Maldoni, R.. "The DaQuinCIS architecture: a platform for exchanging and improving data quality in Cooperative Information Systems", Inform. Syst. 29, 7, pp **(551–582)** **,January** **2004.**

[21] De amiciS, F. and Batini, C, "A methodology for data quality assessment on financial data". Studies Commun. Sci. SCKM, **2004**.

[22] M.J. Carey, L.M. Haas, P.M. Schwarz, M. Arya, W.F. Cody, R.. II,J.H. Williams, and E.L. Wimmers, "Towards heterogeneous multimedia information systems: The Garlic approach", IBMAlmaden Research Center, San Jose, CA, **1996.**

