© 2016 IJSRSET | Volume 2 | Issue 1 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Study of a Detection and Elimination of Data Inconsistency in Data Integration

Dattatray R. Kale, Smita. Y. Aparadh

Department of Computer Engineering, Ashokrao Mane Polytechnic Vathar Tarf Vadgaon, Kolhapur, Maharashta, India

ABSTRACT

Data quality is highly important for running the effective business process. The real world data is spread over the various locations.A collections of these data from the different data sources and presenting the entire collection as a single source is difficult. Data integration involves combining data from numerous dissimilar sources, which are stored using different technologies and present a unified view of the data.Heterogenous and homogenous data is presented at various locations. A big problem in data integration is conflicts occurred into various data sources. Data Inconsistency exists when various and conflicting stories of the same data appear in different places. Data inconsistency shows unreliable information. So in this paper we are presenting the various techniques for finding data inconsistency in data integration.

Keywords: Data Inconsistency, data integration, data source quality.

I. INTRODUCTION

Data quality is recognized as one of the most important problem in data management [1]. Unfortunately the real world data is often dirty that shows the inconsistency is presented in the database. Data integration involves combining data from a number of dissimilar sources, which are stored using various technologies and provide a unified view of the data. Data integration becomes increasingly important in cases of merging systems of two companies or consolidating applications within one company to provide a unified view of the company's data assets. Data integration has been a long standing challenge in the database. Facing the data inconsistency is one of the important challenges in data integration.

A data sources sometimes conflicts with each other. Sources are in different data models or have different schemas within the same data models. A data in the various sources is represented in the different natural languages. A data is presented with different measurements system. There are some differences between the values of data for a single object.

Data inconsistency occurs when different versions of the same data appear in different places. An inconsistency presents when two objects obtained from dissimilar data sources are recognized as versions of each other and a number of the values of their corresponding attributes are different. Data inconsistency can be detected in query processing when a user fetches a query for obtaining a particular value. Data inconsistency creates unreliable information, because it is difficult to determine which version of the information is correct. For example, in data integration system, two student records stored in two different data sources refer to the same student name. The age of the same student name recorded in these two data sources are 12 and 15 respectively. This shows a data inconsistency. If we propose a query in data integration system about the age of given name of student, data inconsistency will be detected from the query result.

A data is spread over the various sites known as distributed data having some benefits and drawbacks. A distributed database system permits applications to access data from local and remote databases. In a homogenous distributed database system, every database is an Oracle database. In a heterogeneous distributed database system, at least one of the databases is a non-Oracle database. Distributed databases use client-server architecture for processing information request. In distributed data at one source data are currently updated whereas other is oldest. Some data come from authoritative sources, whereas other may have uncertain family background.

So in this paper we are studying the use of data source quality criteria to resolve data inconsistency in data integration. We describe the data source quality criteria and studying the various data integration models which satisfying the data sources quality criteria. We are studying the incremental detection of inconsistencies in distributed data.

II. RELATED WORK

There are some techniques for detecting and removing data inconsistency. Several research projects have tackled the detecting and finding the data inconsistency. The first prime importance is to detect the inconsistencies within a centralized as well as distributed database. After the detection of inconsistency we need to remove from centralized as well as distributed data. In data integration based on the content of conflicting data which detects the existence of data inconsistencies and provides users with some additional information on their nature and removes the data inconsistency through the use of probabilistic data. This technique having some advantages in some parts but they have not taken the considerations of data source quality.

There are some research works on data inconsistency solution based on data source properties. These research works can only process data source properties with quantitative values. Current days it is fail to satisfy the needs of business applications. To describe characteristics of data sources better, the qualitative values should be used.

Another one research can only provide fusion strategy instead of some specific algorithms and does not give the illustration of its data inconsistency solution effects.

III. DATA SOURCE QUALITY PRINCIPLES

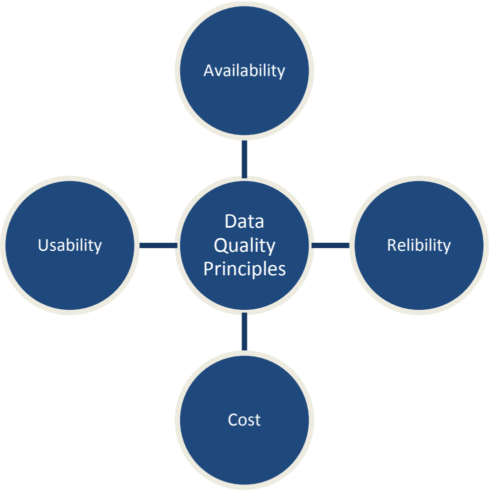
We need to mention the quality principles for obtaining the consistent data. The value of various quality standards is provided by data source itself, some values are obtained from other internet users, and there are also web sites that are used for calculating the values of data source quality principles.

Within an organization, satisfactory data quality is critical to operational and transactional processes and to the reliability of business analytics or business intelligence reporting. Data quality is affected by the way data is entered, stored and managed. Data quality assurance (DQA) is the process of verifying the reliability and effectiveness of data [3].

Maintaining data quality needs going through the data from time to time and scrubbing it. Typically this involves updating it, standardizing it, and eliminating the redundant records to create a single view of the data, even if it is stored in multiple disparate systems. There are a lot of dealer applications on the market to make this job easier.

Business processes depend on timely and accurate information. If information is inaccurate or missing, decisions cannot be made, which may lead to undesired results. In view of the fact that information is built on data we must have imminent in the quality of our data sources. Let's have a look at the different Data Quality aspects. Data Quality and Information Quality can be related to four areas and depend on how these areas are organized and collaborate. In order to evaluate these areas we need to define Data Quality principles.

There is some data source quality principles used for finding the data inconsistency. The main four data quality principles are Availability, Usability, Reliability, and Cost which shows that how much data is available, can we use the available data? Can we trust the available data? And how much cost of the available data.



IV. DATA MODELS FOR DATA INTEGRATION

In this paper we are studying data integration data models. The first data model shown by [12] is mainly focuses on the data sources quality criteria. This data model used classes and class hierarchy, a set of attributes associated with each class and a data source quality criteria vector associated with each attribute for a certain class. These classes contains objects And the attribute value should be atomic value such as a string or integers. There is a one condition that an object can belong to more than one class. It is possible to assert a pair of classes to be put out of place which means that no object can belong to both classes. In this data model two kinds of data schemas are used namely local schema and global schema. The local schemas are provided by data sources in data integration system to describe each data sources local data with local classes and generate local class tree. The global schemas are obtained by integrating the local classes in local schemas and generating global class tree which describes all the data in data integration system with global classes. The another model is Conceptual Data Integration Model which is an implementation free representation of data integration requirements for the proposed system that will serves as a basis of scoping how they are to be satisfied and for project planning purpose in terms of source system analysis, task and duration, and resources. This data model is only necessary to identify the major conceptual processes to fully understand the user's requirements for data integration and plan the next phase. Another data model is named as Logical Extraction Data Integration Models which determines what subject areas will need to be extracted from sources, such as what applications, databases and unstructured sources. Source file formats should be mapped to the attribute level. Once extracted, source data files should be loaded by default to the initial area. Extract integration models consist of two different sub processes.

- Getting the data out of the source system.
- Formatting the data to a subject area file.

Another model is Logical Transform Data Integration Models [6] which identifies at a logical level what transformations that means in terms of calculation; splits, processing, and enrichment are needed to be performed on the extracted data to meet the business intelligence requirements in terms of aggregation, calculations, and structures. Transform types as defined in the transformation processes are determined on the business requirements for conforming, calculating, and aggregating data into enterprise information.

V. DATA INCONSISTENCY SOLUTION

Data inconsistencies can be detected as per the guidance given by [5].Here it uses keys to identify the objects. Key means an attribute or set of attributes to decide objects referring same object in the real world. If the data inconsistency occurred by the local inconsistent set of attributes, it will resolved by the way of selecting the most suitable object of local class provided by the best data source. Let us consider local attribute set {La} where La is an attribute on local class C. The selection of the best data source will use fuzzy multi-attribute decision making approach [9, 11].

The first step for finding the data inconsistency in query result is to obtain Fusion Matrix. Here we introduce triangular fuzzy number [9] to represent the values of qualitative criteria. For qualitative criterion, it can use triangular fuzzy number scaling method which is improved from bipolar scaling method to transform the value of it into triangular fuzzy number. The next step is to construct the fusion decision matrix and compute distance to the positive ideal solution and negative ideal solution for alternatives. For example, a poly object of global class: employee (ID, Name, Age, Salary) may be visualize as by following table.

Table No.1 Example of Polyobject

ID	Name	Age	Salary
6233	Max	38(qv)	5000
6233	Max	35(qv)	4000
6233	Max	40(qv)	9000
6233	Max	35(qv)	7000

Table No.2 The original inconsistent data and decision data.

Value of global attribute ‘Age’	38	35	40	35
Sources Criterias	S1	S2	S3	S4
Time	68	56	28	37
Cost	49	85	76	29

In the Table2, the global attributes “Age” and “Salary” are inconsistent global attributes. ”qv” represents the data source quality criteria vector of local class which provides the corresponding attribute value. For the consistent global,”qv” can be ignored such as “ID” and “Name”. As describe in Table3, “Age” is inconsistent global attribute used as an example for inconsistency solution strategy. By applying the various steps as described above we can find out an inconsistency.

VI. CONCLUSION

In this paper, we have presented a solution for finding out the data inconsistency in data integration. As verified by experiment our strategy has very good performance. In future we will look to improve the accuracy of data inconsistency solution.

VII. REFERENCES

[1] P.Anokhin, “Data Inconsistency detection and resolution in the integration of heterogeneous information sources”, Ph.D, Thesis, School of Information Technology and Engineering, George Mason University, 2001.

[2] S. Agrawal, S. Deb, K. V. M. Naidu, and R. Rastogi, Efficient detection of distributed constraint violations, in ICDE, 2007.or and Second Author.

[3] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C-A.Saita. Declarative Data Cleaning: Language, Model, and Algorithms. In International Conference on Very Large Data Bases, pages 371, 380, 2001.

[4] X.Chai Sayyadin, A. Doan, A.Rosenthal, and L.Seligman,”Analyzing and revising data integration schemas to improve their matchebility” in proceeding of 34th international Conference on very large Data base, 2008, PP, 773-784.

[5] A.Motro and P.Anokhin,”Fusionplex: Resolution of data inconsistency in the integration of heterogeneous data sources” Information Fusion, Vol.7, 2006, pp.176-196.

[6] M.A.H.Andez, S.J.Stolfo, and U.Fayyad,”Real-world data is dirty: Data Cleaning and the merge/purge problem” Data mining and knowledge Discovery, Vol2 1998 pp.9-37.

[7] Y.Papakonstontinou, S.Abiteboul and H.Garcia-Molina,”Object fusion mediator systems” in proceeding of 22nd international conference on very large database, 1996, pp.413-424.

[8] R.Y.Wang and D.M.strong,”Beyond accuracy: what data quality means to do consumers”, Journal of management Information systems, Vol.12, 1996, pp.5-30.

[9] J.M.Benitez, J.C.Martin and C.Roman,”Using fuzzy number for measuring quality of services in hotel industry”Tourisum management, Vol.28, 2007, pp.544-555.

[10] F.E.Uzoka “A fuzzy enhanced multicriteria decision analysis model for evaluating university academics research output”, *Information knowledge systems management*, Vol.7, 2008, pp.273-299.

[11] L.A. Zadeh,”The concept of linguistic variable and its application to approximate reasoning,” Information Science, Vol.8, 1975, pp.199-249.

[12] XIN WANG,LIN-PENG HUANG,XIAO-HUI XU,”A solution for Data Inconsistency in Data Integration”, Journal of Information Science and Engineering 27,681-695(2011).