IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 13, Issue 3 (Jul. - Aug. 2013), PP 68-72
www.iosrjournals.org

Approximating Source Accuracy Using Dublicate Records in Data Integration

[1] Ali Al-Qadri, [2]Ali El-Bastawisi, [3]Osman Hegazi
[1] Department of Information System, , Cairo University, Egypt
[2,3] Professor, Department of Information System, , Cairo University, Egypt

Abstract: *Currently, there are two main basic strategies to resolve conflicts in data integration: Instance-based strategy and metadata-based strategy. However, the two strategies have their limitations and their problems. In a metadata-basedstrategy, the data integration administrator uses some features of source to determine how they may participate to provide a single consistent result for any conflicted object.*

One of the most important featuresin metadata-based strategyis source accuracy, which is concerned with the correctness and precision with which real world data of interest to an application domain is represented correctly in the data store and used for conflict resolution. However, it is difficult and time consuming to calculate accuracy for all data sources and compare data about objects with what is in real world, in this article we propose a new way to approximate sources accuracy using known accuracy of predefined source as a reference. In this paper wepresenthow to determine most appropriate source as a reference using some criteria do not depend on comparing with real objects. After that we demonstrate how to use this reference source and its known accuracy to approximate other sources accuracies using common objects between them and reference source.

Keywords: *Data Integration, Metadata, Source Accuracy, Similarity ratio.*

I. Introduction

Data Integration refers to the problem of combining data residing at autonomous and heterogeneous data sources, and providing users with a unified global schema [1, 2]. It can be defined as the process of combining data residing at different data sources, and providing the client with a unified global view of these data. In addition to provide the client with a single consistent result for every object represented in these data sources. Data integration system allows its users to perceive the entire collection as a single source, query it transparently, and receive a single and unambiguous answer [3]. The sources may conflict with each other on the following three levels: their schema, data representation or data themselves. In this paper we will discuss data inconsistency level in data integration which called in some studies instance-level inconsistency.

The value for some attributescan be provided by more than one source. Several sources compete in filling the result tuple with an attribute value. If all sources provide the same value, that value can be used in the result, but if the values differ, there is a data conflict and resolution function must determine what value shall appear in the result table in global schema [4].

In metadata-based strategy, inconsistencies can be resolved based on sources metadata or in other words based on the qualifications of the data sources.

The concept of data accuracy introduces the idea of how precise, valid and error-free is data: Is data in correspondence with real world? Is data error-free? Are data errors tolerable? Is data precise enough with respect to the user expectations? Is its level of detail adequate for the task on hand? [5].

The most common definitions of data accuracy concern how correct, reliable and error-free are the data [6] and how approximate is a value with respect to the real world [7].

The value of accuracy factor for each attribute in data source is calculated as the distance between the actual value (v) and the database value (v'), after that we can make a normalization for each value alone to 0 or 1 then get whole ratio, or we can take the average of these values for data source or representative sample fromit, as shown in section 4. And this operation must be applied for all sources participating in data integration system. We can observe that it is difficult and time consuming operation; so that we can apply our technique to overcome this challenge by only make this calculation for one source and use common records to approximate other sources accuracies.

The rest of this paper is structured as follows: Section2 show must important metadata; section 3 will discuss how to choose one source as reference source. In section 4, we list all ways used to calculate accuracy for reference source. While section 5 illustrates how to approximate source accuracy using duplicate objects and accuracy of reference source.

II. Source Set of Features

It is common for the data sources being integrated to provide conflicting information about the same entity; so that, the major challenge for data integration is to derive the most complete and accurate integrated records from diverse and sometimes conflicting sources[8].

Each data source in data integration system has a set of features [9, 10]:

-Accuracy: Probabilistic and statistical information that denotes the accuracy of the information or syntactical accuracy of data values that can be checked by comparing data values with reference dictionaries (e.g. address lists, name dictionaries, area zip code lists, and domain related dictionaries such as product or commercial categories lists).

-Timestamp: The time when the information in the source was validated.

-Completeness is the degree to which values of a schema element are present in the schema element instance.

-Internal consistency is the degree to which the values of the attributes of an instance of a schema element satisfy the specific set of semantic rules defined on the schema element; or it is the measuring of intra-source integrity constraints

-trustworthiness: which depends on who is the responsible on data source. Is it confident organization? Is source administrator expert?

-Cost: The time it would take to transmit the information over the network and/or the money to be paid for the information.

-Availability: The probability that at a random moment the information source is available.

-Clearance: The security clearance level needed to access the information.

These set of features can come from different ways, either from the data source themselves like date of update, or through global system or multidatabase administrator which they could assign and maintain and modify meta-data values or scores for local data sources frequently, or by using some websites which dedicated to evaluate other sites.

Most of these features effect on source accuracy, and accuracy reflects some features confidence, so that source accuracy feature considers the most important one.

III. Reference Source Selection

We classify source features into three categories; first category which needs a data store comparisons with what is in real, such as accuracy which this paper concern with; second category includes features that don't depend on comparing with real objects and related with accuracy, such as completeness,timestamp, internal consistency andtrustworthiness; Thirdcategory can be named Quality of Service which depends on the quality of delivering data, such as cost, security, availability and responsiveness.

In this paper, our concern is accuracy which falls under first category, in addition to second category which affects accuracy as follows:

1.1. Completeness and Accuracy

Completeness defined as the degree to that all data relevant to an application domain have been recorded in an information system [11]. It expresses that every fact of the real world is represented in the information system [12]. Completeness can be calculated either by the ratio of entities included to the whole required entities, or by the number of not null values to the whole values for required attribute, which must be reflected on the global schema; the second definition of completeness is used in this research, and it is helpful to make fairness between different data sources by only comparing completeness of similar attributes in all sources.

Some studies [9] define errors or null values in some attributes as value inaccuracy, because that real world data not referenced.

1.2. Internal Consistency and Accuracy

Internal consistency is the degree to which the values of the attributes of an instance of a schema element satisfy the specific set of semantic rules defined on the schema element; or it is the measuring of intra-source integrity constraints.

Internal consistency implies that two or more values do not conflict each other or obey to some correlation between these attributes in the same record. A semantic rule is a constraint that must hold among values of attributes of a schema element, depending on the application domain modeled by the schema element [13].

As an example, if we consider employee with attributes Name, DateOfBirth, Sex, and DateOfHire, some possible semantic rules to be checked are:

- The values of Name and Sex for an instance are internally consistent;If Name attribute is compatible with Sex attribute in used name dictionary.
- The value of DateOfBirth must precede the value of DateOfHire.

In inconsistent data source and according to the description of this feature we can find that employee of 25 old years was born in 1965 or was born in 2020 which can represent inaccurate data source which reflects that internal consistency effects on accuracy.

1.3. Timestamp and Accuracy

Timestamp, uptodateness or currency can be measured by the lag between the time real-world data changes and the time changes are represented in the system. Timestamp can be defined also as the time when the information in the source was validated.

A data value is up-to-date if it is correct in spite of possible discrepancy caused by time-related changes to the correct value; a datum is outdated at time *t* if it is incorrect at *t* but was correct at some time preceding *t* [7].

According to this definition we can observe that non up-to-date data maybe inaccurate data.

1.4. Trustworthiness and Accuracy

Data owned by governmental organization may have more creditability and accuracy than data owned by other types of organizations.

Data collected and reviewed by expert users are more accurate than data collected and reviewed by non-expert users.

We can use these four features to determine which source is eligible to be the reference for rest sources, then accuracy calculated for chosen source as shown in next section.

Data integrator can give equal or non-equal weight for these features to determine which source is more accurate.

Sources with difficulty in finding real objects for comparison operation can be neglected and exceed this step.

IV. Reference Source Accuracy Measurement (The Normal Way)

The accuracy value of data item can be represented as Boolean value (1 for true value and 0 for false), or it can be represented in [0-1] range to calculate confidence or accuracy degree, or it can be represented as numeric value that captures the distance between data value and reference value, and this numeric distance value can be normalized to [0-1] range, these are called metric types.

There are many ways to calculate a global accuracy for a whole relation [5]:

−Ratio: This technique calculates the percentage/ratio of accurate data of the system [14]. This percentage is calculated as the number of accurate data items in the system divided by the total number of data items in the system. The accuracy of data items is expressed with Boolean metrics, i.e. $ai \in \{0, 1\}$, $1 \leq i \leq n$. The accuracy of S is calculated as:

$AccuracyRatio(S) = |\{ai / ai = 1\}| / n$

A generalization can be done for the other types of metrics, considering the number of data items whose accuracy values are greater than a threshold $\Theta, 0 \leq \Theta \leq 1$

$AccuracyRatio(S) = |\{ai / ai \geq \Theta\}| / n$

− Average: This technique calculates the average of the accuracy values of data items. The accuracy of data items can be expressed with any type of metric. The accuracy of S is calculated as:

$AccuracyAvg(S) = (\sum iai) / n$

This technique is the most largely used, for the three types of metrics. Note that if accuracy of data items is Boolean values the aggregated accuracy value coincides with a ratio.

− Average with sensibilities: This technique uses sensibilities to give more or less importance to errors and calculates the average of the sensitized values. Given a sensitivity factor α, $0 \leq \alpha \leq 1$, the accuracy of S is calculated as:

$$AccuracySens(S) = (\sum_i ai^{\alpha}) / n$$

− Weighted average: This technique assigns weights to the data items, giving more or less importance to them. Given a vector of weights W, where *wi* corresponds to the i-th data item, $\sum iwi = 1$, the accuracy of S is calculated as:

$$Accuracyweight(S) = \sum_i wiai$$

Source providers or domain experts can also provide error ratio estimations based on their knowledge/experience with the data.

V. Approximating Source Accuracy

In this section we will approximate sources accuracy depending on accuracy of one source using duplicate (i.e. if there is no duplicate there is no need to use metadata to resolve conflicts) by taking in our minds three assumptions:

1. The data sources are independent (no one copy and paste from the other which common in web sources).
2. "The probability that S1 and S2 provide the same false value is very low with taking in our minds the independent assumption" [15].
3. The duplicated objects between two data sources are representative samples for their sources, so thatit is possible to use more attributes from different tables to increase duplicate records between reference source and other sources if needed. This assumption is justified by some studies that state "The measurement process for accuracy (and for completeness if considered) can be performed on a sample of the database. In the choice of samples, a set of tuples must be selected that are representative of the whole universe and in which the overall size is manageable"[13].

Table 1: common five employees from two sources**.**

	Source 1		Source 2	
Identifier	Name	Age	Name	Age
1234	Ahmed.Kamal	26	Ahmed.Khalifa	26
1987	David.Petter	32	David.Petter	27
2345	Tharwat.Abdu	29	Tharwat.Abdu	29
7654	Ali.Lotfy	30	Aly.Lotfy	30
9121	John.Palin	37	Jon.Palin	37

Example 4.1.Consider the two data sources provide name and age information about the five employees.

5.1. Similarity Ratio Calculation

We will calculate thesimilar values in all duplicate records between source 1 and source 2 using threshold for numerical attributes. Administrator can provide tolerance level (i.e. if the difference between two values <= 2 they considered equal).

With respect to nun-numerical attributes we usedJero-Winkler similarity measurement algorithm. After that we normalize similarity between each two fields to be represented as boolean values (i.e. if similarity>=0.90 they considered equal and represented as 1, else they considered non-equal and represented as 0).

In our example, age attribute has only one conflict withemployee identified by 1987, by using Jero-Winkler algorithm in Name attribute the similarity ratio was 0.82, 1 , 1 , 0.93 and 0.904respectively, so if we normalize with 0.90 threshold we can observe that conflict occur with employee identified by 1234 (i.e. similarity between Ahmed.Kamal and Ahmed.Khalifa is 0.82 and represented as 0).

According to this calculation, similarity ratio will be 0.8 and conflict ratio will be 0.2 in our illustrative example.

We can denote the similarity between source 1 and source 2 as Sim(S1,S2). And we can denote the conflict between them as Conf(S1,S2).

5.2.Accuracy Approximation

If accuracy of source number n denoted by A(Sn), so that accuracy of source 1will be denoted as A(S1) and accuracy of source 2 will be denoted as A(S2) (i.e. source 1 is the reference source in our example and its accuracy must be calculated in normal way calculation with comparing source objects with what is in real world as shown in section 3 and we assume thataccuracy of source 1 is 0.9 in example).

While accuracy of source 2 can be calculated by taking the intersection between accuracy of source 1 which is the reference source, also similarity ratio between two sources did not depend on accuracy of any one of them; and by applying conditional probability, the intersection will be the multiplication of them, so that:

$$A(S2) = A(S1) * \mathrm{Sim}(S1, S2) \qquad (1)$$

In our example:

$$A(S2) = 0.9 * 0.8 = 0.72$$

The two non-similar values in the table may be true or not with respect to source 2, but we cannot decide which is the source that provide true values and which did not, so we will make another sources in data integration system participate in voting about these conflicted values.

This voting depends on asking other sources, if they provide the same objects that cause conflict, either all or part of them.

According to assumption 2 described above, if other source supports source 2 values,non-similar values will be considered true with respect to source 2, and if it supports source 1 it will be considered false with respect to source 2.

So that $A(S2) = 0.72 + 0.2 = 0.92$ If the two conflict values provided by source 2 are true after voting. And $A(S2) = 0.72 + 0 = 0.72$ if the two conflict values provided by source 1 are true after voting. And $A(S2) = 0.72 + 0.1 = 0.72$ if the only one conflict value provided by source 1 are true aftervoting and we will change conflict ratio Conf(S1,S2)=0.1

If there is no another source provides same or part of conflict objects we will calculate probability that make conflict ratio true withrespect to source 2.

In other word if source 2 is accurate in conflicted fields, source 1 is inaccurate in these fields, so if we denote probability that source 2 is accurate in conflict fields as P(Conf):

$$P(Conf) = [1 - A(S1)] * Conf(S1, S2) \quad (2)$$

In our example:

$$P(Conf) = [1 - 0.9] * 0.2 = 0.02$$

The final equation of approximating accuracy of source 2, if there is no voting happens or only partial vote happens:

$$A(S2) = A(S1) * Sim(S1, S2) + [1 - A(S1)] * Conf(S1, S2) \quad (3)$$

In our example:

$$A(S2) = 0.72 + 0.02 = 0.74$$

5.3. Logical analysis

If it is assumed that S1 is absolutely accurate 100% then the accuracy of S2 must be 80% because the conflicted values have no mean and must be false because S1 have no false probability, so if we use our equations can we prove that?

$$A(S2) = (1 * 0.8) + [(1 - 1) * 0.2] = 0.8$$

Then $A(S2) = 0.8$As assumed to be.

The way we use to calculate accuracy for source 2 will be repeated for all sources. After that, approximated accuracies can be used to decide which source is more accurate and which has more priority to be used in conflict resolution operation.

VI. Conclusion And Future Work

In this paper, we have supposed an algorithm to approximate sources accuracies using reference source and taking the advantages of duplicate happen between data sources in data integration.

Other studies deal with duplicate records and conflicted attributes as a problem, but, this paper depends on this duplicate to approximate accuracy to be used in conflict resolution in next phases beside other metadata. In this paper, accuracy of reference source derived from whole data source, in future researcher can apply this study on table or attribute granularity by calculating accuracy for tables and attribute, not for whole data source. In future researcher can show how to apply this approximating for other source feature.

References

[1] Y. Halevy. Answering queries using views: A survey. Very Large Database J., 10(4):270–294, 2001.

[2] R. Hull. Managing semantic heterogeneity in databases: A theoretical perspective. In Proc. of the 16th ACMSIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS'97), 1997.

[3] A. Z. El Qutanny, A. H. Elbastwissy, O. M. Hegazy. A Technique for Mutual Inconsistencies Detection and Resolution in Virtual Data Integration Environment, IEEE ,2010.

[4] F.Nauman and M.Haussler, Declarative data merging with conflict resolution. International Conference on Information Quality. 2002

[5] V.Peralta. Data Freshness and Data Accuracy: A State of the Art. Technical Report. 2006.

[6] R.Wang, Strong. Beyond Accuracy: What data quality means to data consumers. Journal on Management of Information System vol12(4):5-34,1996.

[7] T.Redman. Data Quality for Information Age. Artech House,1996.

[8] B. Zhao, B. Rubinstein, J.Gemmell, J. Han: A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. PVLDB 5(6): 550-561 ,2012.

[9] A. Motro, and P.Anokhin.Fusionplex: Resolution of Data Inconsistencies in the Data Integration of Heterogeneous Information Sources. Information Fusion, 2005.

[10] M. Angeles, M. MacKinnon. Solving Data Inconsistencies and Data Integration with a Data Quality Manager, school of Mathematical and Computer Sciences, Heriot-Watt University, 2004.

[11] M.Gertz,M.TammerOszue, G.Saake, K.Sattler . Report on the dagustuhl seminar: Data Quality on the web. SGMOD Record vol 33(1),2004.

[12] M.Bobrowski, M.Marre, D.Yankelvich. A Softwate Engineering View of Data Quality. 2nd Int. Software Qualitee Week Europe(QWE'98), Brussels, Belgium, 1998.

[13] G.Shanks, B.Corbitt. Understanding Data Quality: Social and Cultural Aspects. In Proc. Of the 10th Australasian conference on Information System, Wellington, New Zealand, 1999.

[14] H.Kon,S.Maknick,M.Siegel. Good Answers from Bad Data: a Data Management Strategy. Working paper 3868-95, Sloan School of Management, Massachusetts Institute of Technology, USA, 1995.

[15] A. Fuxman, E. Fazli, and R. J. Miller. Conquer Efficient management of inconsistent databases. In Proc. of SIGMOD, pages 155–166, Baltimore, MD, 2005.

[16] P.Fendo.General guidelines on information quality:structured, semi-structured and unstructureddata ,FIRB.2003