Information Validity Assessment in Integrating Heterogeneous Data Sources

Peter P. Chen
Computer Science Department
Louisiana State University
Baton Rouge, LA 70803, USA
pchen@lsu.edu

Abstract-- Validity assessment of information from various data sources is a crucial problem in information fusion. This paper identifies several key issues and proposes a framework to solve this problem. It also describes a prototype of a decision-support system to support the estimation of the composite data values from heterogeneous databases with different validity assessment values.

Keywords: information validity, data validity, information fusion, data quality, heterogeneous data, data integration, meta data, decision-support system.

1 Introduction

Information is viewed as an active interaction of knowledge and data. An active paradigm for information management is proposed for investigation of the proposed information services. Under this active paradigm, information services are viewed as behaviors of the information system to be implemented as "intelligent" shareable software modules to provide cooperation among users, applications/agents, and the information system environment.

Intelligent extraction and validation of information from multiple database systems that are of different models (e.g. relational, Object-Oriented, flat files, etc.) are the key prerequisites to efficient knowledge integration for decision support. The extraction and validation capabilities can serve as mediators between the users or applications/agents and the underlying information system in carrying out complex tasks.

Current commercial database systems (e.g. Sybase, Oracle, etc.) and most of the research prototypes focus on providing information for the users through query mechanisms. Our effort attempts to provide an active information management framework, not only allowing applications or users query or extract information from databases of different model (e.g. relational, Object-Oriented, flat files etc.), the active information system cooperates with the users by automatically informing the applications or users of the necessary information whenever critical situations occur. One of the key aspects in our effort is to develop a framework and techniques for information validity assessment.

In many situations (military/civil environments and our daily lives), we encounter the information validity problem in many different forms: the data from different sources are supposed to be the same, but actually, they are different. Here are some examples:

- Three sensors are tracking the same military target, but each is reporting a different location of the target.
- The intelligence reports provide conflicting data. For example, at the end of the World War II, Hitler received conflicting intelligence reports on the location and date of the pending invasion of the European Continent by the Allied Forces.
- Two databases in different locations started with the same data and schema. After a while, these two copies are drifting apart because new data fields are added (and some existing data fields are deleted or changed) for the schema of each database due to local needs. In addition, some of the data in the unchanged data fields could also be different because of maintenance and other errors.
- Some of the employee data (of the same employee) in different databases of the same organization are different.

How can we handle this problem? We think there are at least three things that will be very useful:

- A framework for analysis
- Computational formulae or algorithms for conflict resolution.

* This research is supported in part by AFRL and ONR funds via contract #F3060299MV037.
• A software program that help people make decisions.

We will describe each of these three subjects in the following sections.

2 A Preliminary Framework for Information Validity Assessment

Here, we will try to outline several major steps of framework for the information validity Assessment.

2.1. Factors of Information Validity Assessment Value

First, we need to know the “factors” of Information validity (InfV) Assessment.

Each data source will be given an InfV value. However, InfV is a composition of several factors including the following factors:

- **Reliability of the hardware and software that the database resides**: Collecting the past operational data of the hardware and software malfunction frequencies can be useful in assessing this reliability value in the past time periods. The reliability value is between "0" and "1."

- **Freshness of the data**: If the data was just updated, the freshness factor could be 1.0. If the data was updated one year or more ago, the freshness could be "0". We can define a function of values between "0" and "1" for any "age" of data between "one year" and "0".

- **Believability of the data**: how believable of the data depends on where the data came from. For example, if the data comes from a very believable source, it will be certainly more believable than the data comes from an unreliable source. The believability factor has value between "0" and "1."

3 Computational formulae or algorithms for conflict resolution

3.1. The Need for Conflict Resolution Algorithms and a Collection of Some Formulae

If we derive 3 sets of values from three different sources for the same query, then the question is which one is correct. Or, how do we derive an Inv(V) value from the sets of values derived? One approach to solve this problem is to associate a value between 0 and 1 to the BASE data value (instead of assuming them to be always 1 as in the case of traditional databases). The InfV values can be stored either in the database or the meta database based on the original conceptual model or the conceptual model resulting from reverse engineering/modeling.

Typically, the value associated with the base data depends on the source of data acquisition, update frequency, credibility of data, the confidence in the system from which the data is obtained or the confidence in the instrument from which the data is acquired, etc. Once a measure is associated with the base data, they can be combined (or new values can be derived) in a number of ways.

The following is a collection of some simpler formulae or algorithms.

1. Take the max of the values
2. Take the min of the values
3. Take the average of the values
4. Take consensus or majority vote
5. Discard values below or above a threshold and apply the above to the resulting values.
6. Arbitrarily pick one
7. Apply a function on the probability to compute the new values
8. Dynamically collecting probabilities by the system

1.1 A Proposed Algorithm

In the following, we are going to describe a new algorithm to derive a “composite” value.

Algorithm 1 (for two data sources):

Let $V_1, V_2$ be the data values of the same data element from two different sources, and Let $C_1, C_2$ the data validity...
assessments of the two data values. Let \( V^* \), \( C^* \) be the estimated data value and the associated data validity assessment value based on these two given data values and their associated data validity assessment value, and they can be derived by the following formula:

\[
V^* = V_1 * (C_1 / (C_1 + C_2)) + V_2 * (C_2 / (C_1 + C_2))
\]

if \( C_1 + C_2 \neq 0 \),

\[
(V_1 + V_2) / 2 \quad \text{if } C_1 = C_2 = 0. \quad (1)
\]

\[
C^* = C_1 * (C_1 / (C_1 + C_2)) + C_2 * (C_2 / (C_1 + C_2))
\]

if \( C_1 + C_2 \neq 0 \),

\[
(C_1 + C_2) / 2 \quad \text{if } C_1 = C_2 = 0. \quad (2)
\]

Example #1:
Given these values: \( V_1 = 10 \), \( V_2 = 100 \), \( C_1 = 0.6 \), \( C_2 = 0.8 \), the values of \( V^* \) and \( C^* \) can be derived as follows:

\[
V^* = V_1 * (C_1 / (C_1 + C_2)) + V_2 * (C_2 / (C_1 + C_2))
\]

\[
= 10 * (0.6/0.6+0.8)) + 100 * (0.8 /0.6+0.8))
\]

\[
= 4.29 + 57.14
\]

\[
= 61.43
\]

\[
C^* = C_1 * (C_1 / (C_1 + C_2)) + C_2 * (C_2 / (C_1 + C_2))
\]

\[
= 0.6 * (0.6/0.6+0.8)) + 0.8 * (0.8 /0.6+0.8))
\]

\[
= 0.2571 + 0.4571
\]

\[
= 0.7142
\]

Algorithm 1.1. (for "n" data sources):
Let \( V_1, V_2, \ldots, V_n \) (\( n > 2 \)) be the data values of the same data element from "n" different sources, and Let \( C_1, C_2, \ldots, C_n \) the data validity assessment value of the "n" data values. Let \( V^* \), \( C^* \) be the estimated data value and the associated data validity assessment value based on these "n" given data values and their associated data validity assessment value, and they can be derived by the following formula:

\[
V^* = V_1 * (C_1 / (C_1 + C_2 + \ldots + C_n)) + V_2 * (C_2 / (C_1 + C_2 + \ldots + C_n)) + \ldots + V_n * (C_n / (C_1 + C_2 + \ldots + C_n))
\]

if \( C_1 + C_2 + \ldots + C_n \neq 0 \),

\[
\frac{V_1 + V_2 + \ldots + V_n}{n} \quad \text{if } C_1 = C_2 = \ldots = C_n = 0. \quad (3)
\]

\[
C^* = C_1 * (C_1 / (C_1 + C_2 + \ldots + C_n)) + C_2 * (C_2 / (C_1 + C_2 + \ldots + C_n)) + \ldots + C_n * (C_n / (C_1 + C_2 + \ldots + C_n))
\]

if \( C_1 + C_2 + \ldots + C_n \neq 0 \),

\[
\frac{C_1 + C_2 + \ldots + C_n}{n} \quad \text{if } C_1 = C_2 = \ldots = C_n = 0. \quad (4)
\]

4 A Decision-Support System for Information Conflict Resolution

A decision-support system will be very useful to the persons who need to make decisions based on conflicting data, each with a (possibly) different InfV value. A prototype has been implemented using Java for testing the results on two Oracle databases located in two nodes of a network (one served as a server and the other served as a client). The development language used is Java™ 1.2.2. The operating system used on the server is Windows NT™ 4.0. The database management system used is Oracle 8i. We use Microsoft ODBC Administrator and Net8 Configuration Assistant to set up the oracle client services. We have Oracle 8i client installed on the client workstation running the Java application.

One of the screens of the software prototype is shown in Figure 1. The user can query the data, and the system displays the data from two databases simultaneously. The user can then ask the system’s assistance in selecting one of the algorithms for making decision.

Let us explain the prototype system in more detail. There are two test databases located in two nodes in a local area network. The server station is running under Windows NT, and the client station is running under Windows 98. The test databases are implemented using Oracle 8i. The database schema is shown in the Appendix; it contains some data fields that may be needed in a military application. Both databases have the same schema, but the data are not completely the same. For example, the employee age in one database could be “34” while the age of the same employee in the other database could be “36”.

The user can input SQL query into the blank “input area” at the top of the user screen. The system will respond and fill up the values in the blank areas in the middle of the screen and information validity (InfV) values extracted from both test databases. For example, from Test1 database, the “age” value of a particular employee could be “34” and the InfV value could be “0.8”, while from Test2 database the “age” value of the same employee could be “36” and the InfV value could be “0.6”. The user has a choice of algorithms listed in the list side of the screen. If the user needs explanation of a particular algorithm, the user can click on the buttons on the right-hand side for explanations.”

After the user clicks (chooses) a particular algorithm, the system will calculate the composite data value and the composite InfV value based on the selected algorithm and display the composite values in the blank areas in the middle of the screen. After the user satisfies with the new
composite values, the user can confirm the choice by selecting the correct buttons in the lower right corner of the user screen. The user also has choices of picking one particular data value from one of the two databases. In addition, the user has the choice of overwriting both database values and all algorithms by inputting the data value he/she thinks is correct.

It is possible to use qualitative terms (such as “high,” “medium,” and “low,”) to express InfV values instead of numeric values. However, in order to calculate the composite value of InfV value of the composite data value, we need to translate the qualitative value into numeric values. In Table 3 of the test databases (see Appendix), we keep track of the numeric equivalent values (actually, the range of values and the likely values) of each qualitative terms. When the user asks for explanation of qualitative terms, the system will display the explanation in a pop-up screen (Figure 2) so that the user can get an idea of what the qualitative term means.

5 Conclusions and Future Directions

We have outlined some of the important steps in a framework for information validity assessment. We have also described some algorithms for conflict resolution. We have implemented a prototype of a decision support system for helping people to make decisions under conflicting data situations.

Further research work can be done in the near future. For example, we may consider these extensions: adding a meta-data model, incorporating active modeling, and applying reverse engineering concepts and techniques into the framework. In meta-data modeling and active data modeling, we may consider the use of the Entity-
Relationship (ER) model [1] or one of its extensions. In terms of applications of the framework and techniques mentioned in this paper, one of such application is to assist the identification of the culprits during or after the information attacks in cyberspace [2,3]. A military application is to incorporate the techniques in the architecture, design, and implementation of Joint Battle Space proposed by the Air Force Science Advisory Board. In addition, the software prototype can be extended to accommodate 3 databases or more.

6 Acknowledgment

The author would like to thank Leah Wong and Commander Cory Frank of Navy SPAWAR Systems Center for their comments and suggestions on this research project. The author would also like to thank Ray Liuzzi and Joe Giordano of Air Force Research Laboratory/Information Directorate for their support and encouragement.

References


The testing databases have several tables. One of them (Table 1) looks like the following:

<table>
<thead>
<tr>
<th>EVENT_TYPE</th>
<th>NUMBER(3),</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME</td>
<td>DATE,</td>
</tr>
<tr>
<td>TRACK_NUMBER</td>
<td>NUMBER(10) NOT NULL,</td>
</tr>
<tr>
<td>PLATFORM</td>
<td>NUMBER(3),</td>
</tr>
<tr>
<td>COURSE</td>
<td>NUMBER(5),</td>
</tr>
<tr>
<td>SPEED</td>
<td>NUMBER(5),</td>
</tr>
<tr>
<td>ALT_DEPTH</td>
<td>NUMBER(15),</td>
</tr>
<tr>
<td>RANGE</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>BEARING</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>CALL</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>SIGN</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>MODE_1</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>MODE_2</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>MODE_3</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>MISSION</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>NATIONALITY</td>
<td>CHAR(20),</td>
</tr>
<tr>
<td>SOURCE</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>ID_SOURCE</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>TRACK_TYPE</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>ID</td>
<td>NUMBER(10) NOT NULL,</td>
</tr>
<tr>
<td>CATEGORY</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>SYMBOLOGY</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>LAT_N_S</td>
<td>CHAR(2),</td>
</tr>
<tr>
<td>LAT_DEGREE</td>
<td>NUMBER(2),</td>
</tr>
<tr>
<td>LAT_MINUTE</td>
<td>NUMBER(2),</td>
</tr>
<tr>
<td>LAT_SECOND</td>
<td>NUMBER(10),</td>
</tr>
<tr>
<td>LON_E_W</td>
<td>CHAR(2),</td>
</tr>
<tr>
<td>LON_DEGREE</td>
<td>NUMBER(2),</td>
</tr>
<tr>
<td>LON_MINUTE</td>
<td>NUMBER(2),</td>
</tr>
<tr>
<td>LON_SECOND</td>
<td>NUMBER(10)</td>
</tr>
</tbody>
</table>

The 2nd table contains the meta data. This table describes the InfV values of each column of a particular table (for example, Table 1):

- TBL_NAME VARCHAR2(30),
- COL_NAME VARCHAR2(30),
- VALIDITY VARCHAR2(20),
- PRIMARY KEY (TBL_NAME, COL_NAME)

The third table (Table #3) contains the interpretation of some qualitative terms such as “high,” “medium,” and “low”. The schema of Table is shown below:

- QUALITATIVE-TERM VARCHAR2(10),
- PERCENT_LOW NUMBER(4),
- PERCENT_HIGH NUMBER(4),
- PERCENT_ESTIMATION NUMBER(4),
- PRIMARY KEY (QUALITATIVE-TERM)