FUZZY ALGORITHM FOR SELECTION OF RELIABLE INFORMATION OF DATABASES FROM A COLLECTION OF DIFFERENT SOURCES

Jing Yu
jxy02@comp.uark.edu
Department of Computer System Engineering
University of Arkansas, Fayetteville, AR 72701

Yong-qing Li
liy@mail.ecu.edu
Department of Physics, East Carolina University
Greenville, NC 27858

David Wessels
wesselsd@mala.bc.ca
Department of Computing Science
Malaspina University College, Nanaimo, BC

Abstract
With the rapid development of the Internet, there exist multiple data sources on the World-Wide Web that users can access. Selection of reliable information from these different sources becomes increasingly important. The aim of this paper is to develop source-selection algorithms for selecting reliable information from a collection of different sources for database applications. Fuzzy theory and probability theory are used for modeling the uncertainty of selection and to retrieve the ambiguous data from different sources. A computation prototype is implemented for testing the effectiveness of our fuzzy source-selection algorithms.

1. Introduction

The Internet has changed many aspects of our concepts in data storage and retrieval in the age of the information super highway. Today, if we want to search a piece of information through the World-Wide-Web (WWW), hundreds of sources (i.e., web sites) that can provide the same information may appear. For example, suppose we want to know the share value of a company in stock market at a given time: there are many sources on the WWW (such as stock trading agents and newspaper agents) providing the real-time stock data for this company. Unfortunately, the reliability of the data from different sources may vary. Different sources may provide different share values for the same company although only one value is true. The reason for this uncertainty is that some sources may have updated the data but some may be delayed and some may have mistakes in data collection and processing. Facing so many data sources, users may ask which sources provide the most reliable data. Thus, the selection of reliable information from different sources becomes increasingly important. Data users need to find some way to ensure the currently available data source has the most accurate information if the data in the databases are collected from multiple sources.

Data quality and data assurance techniques have been of interest in a broad area [1-4]. We are going to confine our discussion in a database system. In general, though-out

this paper we assume that we are maintaining a database, in which we track a large population of items and each item may be associated with a large set of attributes. There are many different sources for the information in the database and these multiple sources may provide information on any given attribute for any item. The reliability of the data sources may change in the following ways: (1) the reliability of any source may vary over time. Some sources may provide reliable information at a given time but other sources may at the next time; (2) some sources may be more reliable for certain items; (3) some sources may be more reliable for certain attributes. Our objective is, for each attribute of each item, to try to determine which source currently gives the most accurate information.

In this paper, we will develop source-selection algorithms for selecting reliable information from a collection of different sources for database application. Our approach is to describe the uncertainty in the retrieved data due to the randomness from multiple data sources by using probability theory and to describe the uncertainty due to the fuzziness in the definition of reliability or accuracy by using fuzzy set theory. We then develop source-selection algorithms for selecting the most reliable source for a given attribute of a given item or for a whole item in the cases that (1) the reliabilities of multiple sources for any items in the database are unknown, and (2) the reliabilities for parts of items are prior-known. We also implement a computational prototype for testing the effectiveness of our source-selection algorithms.

2. Modeling fuzziness for reliable data

Assume that we are working on a relational database. Let $R(U)=R(A_1, A_2, …., A_n)$ be a relation scheme on domains $D_1, D_2, …., D_n$, where $U=\{A_1, A_2, …., A_n)$ is the set of all attributes $A_1, A_2, …., A_n$. By definition, each domain D_j is a set of elements of the i-th column in the relation (i.e., a table) and each attribute A_j is the name of a field (i.e., column of the table) played by domain D_j, where j=1,2,….,n. A tuple $t=\{t(A_1), t(A_2), …., t(A_n)\}$ in a relation represents a collection of relational data for an individual item, called a record. A relation R is then described as a table as following.

R

A_1	A_2	…	A_j	…	A_n
		…		…	
$t(A_1)$	$t(A_2)$	…	$t(A_j)$	…	$t(A_n)$
		…		…	

Mathematically, the data value of a given item t_i in the given field associated with the attribute A_j is represented as $R_{ij} = t_i(A_j)$. For an m×n table, the indices are taken as i=1, 2, …, m and j=1, 2, ….., n, where m is the number of rows (i.e. items) in the table, called the cardinality, and n is the number of columns (i.e, fields) in the table, called the degree.

Now, we introduce another index s to indicate the sources of the data. For a data element R_{ij} of the table, it can come from different data sources and, thus, may have different values given by $R_{ij}(s) = t_i(A_j,s)$. Assume that k is the number of the data sources, then, s=1, 2, …, k. So, the data values of an item from a given data source is represented by

$$t(s) = \{t(A_{1,} s), t(A_2, s), … t (A_{n,} s)\}. \quad (1)$$

It is somehow ambiguous to state that the data value $R_{ij}(s)$ (= $t_i(A_j,s)$) of item t_i for attribute A_j from source s is "reliable" or "accurate". In order to model the uncertainty due to the fuzziness in the definition of reliability or accuracy, we use fuzzy set theory [5]. The data value $R_{ij}(s)$ from a given source may be different from the "correct" value R_{ij} for the given attribute of the given item. We define a correlation function to describe the "closeness" between the source data $R_{ij}(s)$ and the "correct" data R_{ij}. This function is known as fuzzication function for quantifying the reliability of the source for the given attribute of the given item

$$f\left(R_{ij}(s)\right) = \exp\left(-\alpha\left|R_{ij}(s) - R_{ij}\right|^2\right), \tag{2}$$

where α is a constant to be chosen. If $R_{ij}(s) = R_{ij}$, we have $f(R_{ij}(s))=1$. In this case, we say that the source s is reliable for providing the data value for the attribute A_j of the item t_i. On the other hand, if the difference $|R_{ij}(s)- R_{ij}|$ is very large so that $f(R_{ij}(s))=0$, then, we say that the source s is not reliable for this data value. Therefore, $f(R_{ij}(s))$ can be used as the characteristic function of "reliability".

3. Modeling randomness for multiple data sources

The derivation of the data values from the "correct" value is randomly distributed over different data sources. Mathematically, the probability of this derivation satisfies Gaussian distribution if the number k of the sources is a large number and there are many reasons causing the data derivation for each source. So, we use the Gaussian statistics to deal with the randomness for multiple data sources as follow. The average of a data value is given by

$$\left\langle R_{ij} \right\rangle = \frac{1}{k}\sum_{s=1}^{k} R_{ij}(s). \tag{3}$$

The variance of the data value is given by

$$\sigma_{ij}{}^2 = \left\langle R_{ij}{}^2 \right\rangle - \left\langle R_{ij} \right\rangle^2 = \frac{1}{k}\sum_{s=1}^{k}\left|R_{ij}(s) - \left\langle R_{ij} \right\rangle\right|^2. \tag{4}$$

The probability that the source *s* provides a random data value $R_{ij}(s)$ is then given by

$$P_N\left(R_{ij}(s)\right) = \frac{1}{\sqrt{2\pi}\sigma_{ij}}\exp\left[\frac{\left|R_{ij}(s) - \left\langle R_{ij} \right\rangle\right|^2}{2\sigma_{ij}^2}\right]. \tag{5}$$

4. Source-selection algorithms

4.1 Algorithm for selecting an individual element without prior known information

In practice, we usually do not know the "correct" data value for each given attribute of the given item. This means that the "correct" data R_{ij} in Eq. (2) is unknown. We also completely do not know the "correctness" of the data values of the other records from each source. In this case, it is reasonable for the users to use the average data value $<R_{ij}>$ as the "correct" value for the first-order approximation because the data values provided by different sources are randomly distributed (with a Gaussian distribution) around the "correct" value. Comparing Eq.(2) and Gaussian distribution in Eq.(5), one may choose α as σ_{ij}^2. Therefore, the fuzziness function for quantifying the reliability of the source can be chosen as

$$f\left(R_{ij}(s)\right)=\exp\left(-\frac{1}{2\sigma_{ij}^{2}}\left|R_{ij}(s)-\left\langle R_{ij}\right\rangle\right|^{2}\right). \quad (6)$$

However, one can choose a different α in Eq.(2) for quantifying the fuzziness function, i.e., $\alpha=\beta/<R_{ij}>^2$ with β a constant. Therefore, for an individual attribute value, $R_{ij}=t_i(A_j)$, we have the reliability characteristic function $f(R_{ij}(s))$ to quantify the question of "which source is more reliable for a data value $t_i(A_j)$".

One may develop a genetic algorithm for selecting an individual data value based on the fuzziness function calculated from Eq. (6). Obviously, $f(R_{ij}(s))$ in Eq. (6) gives the trust factor that a given data $R_{ij}(s)$ from source *s* is reliable comparing to the "correct" value R_{ij}. This trust factor can be used to form a genetic algorithm for source-selection. The data users may believe that a closer data value to the "correct" value R_{ij} than the first-order average of $< R_{ij}>$ is given by the second-order approximation

$$\left\langle R_{ij}^{(2)}\right\rangle=\frac{1}{W}\sum_{s=1}^{k}f\left(R_{ij}(s)\right)R_{ij}(s). \quad (7)$$

This weighted average value can be used as the "correct" value. The new variance is then given by

$$\left(\sigma_{ij}^{(2)}\right)^{2}=\frac{1}{W}\sum_{s=1}^{k}f\left(R_{ij}(s)\right)\left|R_{ij}(s)-\left\langle R_{ij}\right\rangle\right|^{2}, \quad (8)$$

where the normalized factor is given by

$$W=\sum_{s=1}^{k}f\left(R_{ij}(s)\right). \quad (9)$$

By using Gaussian distribution, the new source-selection fuzziness function is given by

$$f^{(2)}\left(R_{ij}(s)\right)=\exp\left(-\frac{1}{2\left|\sigma_{ij}^{(2)}\right|^{2}}\left|R_{ij}(s)-\left\langle R_{ij}^{(2)}\right\rangle\right|^{2}\right). \quad (10)$$

Again, the higher the $f^{(2)}(R_{ij}(s))$, the more reliable the source *s* for the given data value R_{ij}.

A genetic algorithm can be further obtained by replacing

$$\left\langle R_{ij}^{(2)}\right\rangle\rightarrow\left\langle R_{ij}\right\rangle,\ \sigma_{ij}^{(2)}\rightarrow\sigma_{ij},$$
$$f^{(2)}\left(R_{ij}(s)\right)\rightarrow f\left(R_{ij}(s)\right). \quad (11)$$

4.2 Algorithm for selecting a record without prior known information

Similarly, we can also form a source-selection approach to answer the question of "which source is more reliable to provide a record set (tuple), $t_i=\{t_i(A_1), t_i(A_2), \ldots., t_i(A_n)\}$."

For the first-order approximation, we use the average record set of an item as the "correct" value set of the item,

$$\left\langle t_i\right\rangle=\left\{\left\langle R_{i1}\right\rangle,\left\langle R_{i2}\right\rangle,.....,\left\langle R_{in}\right\rangle\right\}, \quad (12)$$

where

$$\left\langle R_{ij}\right\rangle=\frac{1}{k}\sum_{s=1}^{k}R_{ij}(s) \quad \text{for j=1,2, ..., n.} \quad (13)$$

The distance d(s) between the data set $t_i(s)=\{t_i(A_1,s), t_i(A_2,s), \ldots., t_i(A_n,s)\}$ from source s and the "correct" set Eq.(12) is defined as

$$|d(s)|^2 \equiv \left(t_i(s) - \langle t_i \rangle\right)^2 = \sum_{j=1}^{n} \left| R_{ij}(s) - \left\langle R_{ij} \right\rangle \right|^2 . \tag{14}$$

The average square distance is given by

$$\sigma_i^{\,2} \equiv \frac{1}{k}\sum_{s=1}^{k} |d(s)|^2 = \frac{1}{k}\sum_{s=1}^{k} \left(\sum_{j=1}^{n} \left| R_{ij}(s) - \left\langle R_{ij} \right\rangle \right|^2 \right). \tag{15}$$

Similarly, we define a characteristic function for data source selection as

$$f\left(t_i(s)\right) = \exp\left(-\alpha \sum_{j=1}^{n} \left| R_{ij}(s) - \left\langle R_{ij} \right\rangle \right|^2 \right). \tag{16}$$

One can choose $\alpha=1/\sigma_i^2$ or the other values. Again, the larger the $f(t_i(s))$, the more reliable the source *s* for the given data value set of $t_i=\{t_i(A_1), t_i(A_2), \ldots, t_i(A_n)\}$.

4.3 Source-selection algorithm with partly known information

In some cases, we know the "correct" values for some items in the database and we want to select the most reliable for an attribute or all attributes of an unknown item. One example of the situation is that we are adding a new record into our database while the other records are well known. In this case, we can retrieve all records from each source and then compare them with the known records to evaluate the trust factor for each source with a fuzziness function

$$f(s) = \exp\left(-\alpha \sum_{i=1}^{m} \sum_{j=1}^{n} \left| R_{ij}(s) - \left\langle R_{ij} \right\rangle \right|^2 \right). \tag{17}$$

Now, for the retrieved data value $R_{pj}(s)$ of a new record $t_p=\{t_p(A_1), t_p(A_2), \ldots, t_p(A_n)\}$ from source s, we can construct a characteristic function to quantify its reliability as

$$f\left(R_{pj}(s)\right) = \exp\left(-\alpha \left| R_{pj}(s) - \left\langle R_{pj}^{(w)} \right\rangle \right|^2 \right), \tag{18}$$

where

$$\left\langle R_{pj}^{(w)} \right\rangle = \frac{1}{W} \sum_{s=1}^{k} f\left(R_{pj}(s)\right) R_{pj}(s), \tag{19}$$

and

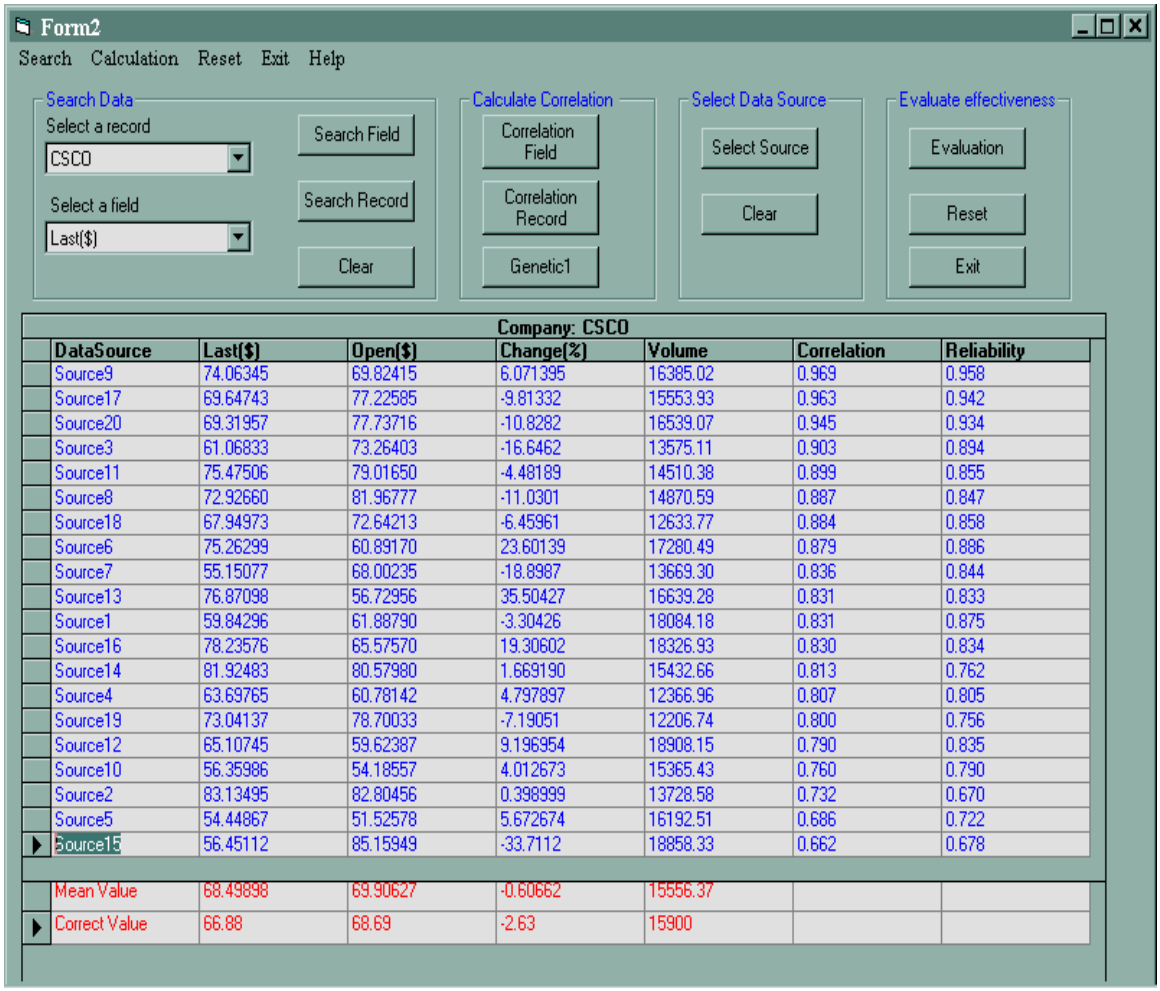
$$W = \sum_{s=1}^{k} f\left(R_{pj}(s)\right). \tag{20}$$

Here $\left\langle R_{pj}^{(w)} \right\rangle$ is the average data value with a weight of $f(R_{pj}(s))$ and W is a normalizing factor.

5. Prototype for testing

We have implemented a computational prototype for testing the effectiveness of our source-selection algorithms. As shown in Fig.1, our program allows search data values of a given record (such as CSCO in stock market) or a given field (such as Last value of CSCO) from a number of sources (20 sources in this example). The data values for each attribute provided from these sources are randomly generated around the "correct value"

of the attribute to model the uncertainty of sources. Four attributes (Last, Open, Change, and Volume) are associated with a record. The program allows us to calculate the mean value for each field over the 20 retrieved data values and allows us to calculate the fuzziness function or correlation function defined in Eq. (16) for each source. Based on this correlation function, users can select the most reliable source. From the calculated result in Fig.1, source 9 is said to provide the most accurate data. As a test for the effectiveness, we use the "correct" values to calculate the "closeness" between the retrieved data values for the searched record and the "correct" values. This "closeness" is shown as a reliability function, which is defined in Eq.(16) but replacing the mean value $<R_{ij}>$ with its corresponding "correct" value shown in the bottom row. The sort result based on the correlation function is very close to that based on the reliability function. This shows that our source-selection algorithms are very effective.



DataSource	Last($)	Open($)	Change(%)	Volume	Correlation	Reliability
Source9	74.06345	69.82415	6.071395	16385.02	0.969	0.958
Source17	69.64743	77.22585	-9.81332	15553.93	0.963	0.942
Source20	69.31957	77.73716	-10.8282	16539.07	0.945	0.934
Source3	61.06833	73.26403	-16.6462	13575.11	0.903	0.894
Source11	75.47506	79.01650	-4.48189	14510.38	0.899	0.855
Source8	72.92660	81.96777	-11.0301	14870.59	0.887	0.847
Source18	67.94973	72.64213	-6.45961	12633.77	0.884	0.858
Source6	75.26299	60.89170	23.60139	17280.49	0.879	0.886
Source7	55.15077	68.00235	-18.8987	13669.30	0.836	0.844
Source13	76.87098	56.72956	35.50427	16639.28	0.831	0.833
Source1	59.84296	61.88790	-3.30426	18084.18	0.831	0.875
Source16	78.23576	65.57570	19.30602	18326.93	0.830	0.834
Source14	81.92483	80.57980	1.669190	15432.66	0.813	0.762
Source4	63.69765	60.78142	4.797897	12366.96	0.807	0.805
Source19	73.04137	78.70033	-7.19051	12206.74	0.800	0.756
Source12	65.10745	59.62387	9.196954	18908.15	0.790	0.835
Source10	56.35986	54.18557	4.012673	15365.43	0.760	0.790
Source2	83.13495	82.80456	0.398999	13728.58	0.732	0.670
Source5	54.44867	51.52578	5.672674	16192.51	0.686	0.722
Source15	56.45112	85.15949	-33.7112	18858.33	0.662	0.678
Mean Value	68.49898	69.90627	-0.60662	15556.37		
Correct Value	66.88	68.69	-2.63	15900		

Figure 1. Prototype for testing

Bibliography

1. R.Y. Wang and D. Strong, Beyond accuracy: what data quality means to data consumers, *J. Manage. Info. Syst.* vol. 12, no.4 (1996), pp.5-34.
2. Becker, S., "A Practical Perspective on Data Quality Issues," *Journal of Database Management*, vol.9, (1998), pp.35-37.

3. R.Y. Wang, V.C. Storey, and C.P. Firth, A framework for analysis of data quality research. *IEEE Trans. Know. Data Eng.* vol.7, no.4 (1995), pp.623-640.
4. Janta-Polczynski, Martin Roventa, Eugene "Fuzzy measures for data quality", *Annu. Conf. North Am Fuzzy Inf Process Soc NAFIPS*, p 398-402, 1999.
5. T. Terano, K. Asai, and M. sugeno, *Fuzzy Systems Theory and Its Applications*, (Academic Press, Inc., Boston, 1991).
6. Guoqing Chen, *Fuzzy Logic In Data Modeling: Semantics, Constraints, and Database Design*, (Kluwer Academic Publishers, Boston,1998).