Reducing of Inconsistent Data Using Fuzzy Multi Attribute Decision Making for Accessing Data from Database

Mohd Kamir Yusof, M. Nordin M Rahman and Atiqah Azlan

Fakulti Informatik, Universiti Sultan Zainal Abidin, Terengganu, Malaysia
mohdkamir@unisza.edu.my

Abstract

Inconsistent data in database occurs due to increasing number of data. A suitable technique is needed to reduce inconsistent data from database. In this paper, fuzzy multi attribute decision making was chosen to reduce inconsistent data from database. This technique contains 5 steps which are deriving quality vector, scale the quality matrix, compute weighted Euclidean distance and select final alternative. An application was developed using java and oracle technology. Sample data was selected for experiments purposes. The result indicates fuzzy multi attribute decision making is a suitable technique in reducing inconsistent data from database. Algorithm in fuzzy multi attribute decision making is able to find out a correct object.

Keywords: *Fuzzy multi attributes decision making, inconsistent data, quality data*

1. Introduction

Database is referring to organized collection of data in digital form. One of the software can be used to manage and maintain a database is Database Management System (DBMS). DBMS is tool for web administrator to insert data, update data and delete data or other operations into database. The important component is database issue is data quality. The problem with data quality is dirty data. The dirty data can be divided into three which are data redundancy, data incomplete and data inconsistency. However this research is focuses on data inconsistency. Data inconsistency exists when two objects (or tuples in the relational model) obtained from different data sources (database) are identified as versions of each other and some of the values of their corresponding attributes differ [2]. This problem happened because of data were stored in database without no standardized format or spelling. The inconsistent data can only detected when processing query involving that particular data. Data inconsistent usually lead to the data redundancy as the query process could not detect the pattern or similarities of the inconsistent data even though they refer to the same thing. Query process will assume that the data were individually different thus causing the same data to be called only with different format or spelling. Besides that, the inconsistent data usually will become unreliable data information as these data were represented in different format. Therefore, without specific detail there is no way for user to identify which representation of data is the correct one. Based on problem statement above, a good or suitable mechanism or technique is needed to reduce inconsistent data during accessing data from database. Several techniques in reducing data inconsistent were described in Section 2.

2. Literature Review

Some of the current techniques in reducing inconsistent data have been studied. Four techniques have been identified which are rough set, logic analysis of inconsistent data

(LAID), fuzzy multi attribute decision making (FMDAM) and functional dependencies (FD) of corresponding relation variable [1]. Rough set theory was proposed by Pawlak in 1982 as a mathematical tool to deal with vagueness and uncertainty in the classification of objects in a set as mentioned in his research [2]. In this theory provides functionality to generate analyze and optimize set of decision rules obtained from data table. Two concepts in rough set theory are lower approximation and upper approximation. The concept of lower and upper approximation can be used to deal with inconsistent object that probably or definitely belong to the set [3, 4]. Differ with second theory, LAID is a method developed based on two existing theories, Rough Set and Logical Data Analysis (LAD). The purpose of this development is to improve the existing theories which are to solve inconsistency created by the process of how sample was developed, that allowed a respondent to belong to more than one class [8]. Meanwhile, FMDAM technique is focuses on inconsistency at data value level that exists when two or more objects obtained from different data sources are identified similar to each other [2]. These type of inconsistency can only be detected when user request for query. In order to resolve the inconsistent issues, data source quality criteria must be first identified. In FD theory, association rule finding algorithm was applied. Association rule is based on how often set of items occur together. This theory focuses on supplying appropriate minimum support based on target database size. Four steps in this theory, 1) select a functional dependency for data inconsistency check, 2) run association rule algorithm for the attributes in the given FD with parameter of minimum support of 1, 3) generate rules for the right hand side of the FD and 4) find inconsistent data with association rule with confidence less than 100% [6, 9]. In Table 1 below, the advantages, disadvantages and critics was analyzed.

Table 1. Analysis about Rough Set, LAID, FMDAM, FD [1]

Techniques	Advantages	Disadvantages	Critics
Rough Set	Able to obtain core and several possible reduction after getting consistent universe -Support many classes and different nominal attribute values	Do not exclude or correct inconsistent data. Allow output discordant decision rules as first and second rules making it difficult for user to interpret the result	More suitable to be implementation in data reduction and data uncertainty process that involve many classes and attribute values
Logic Analysis of Data	Reduce number of attribute in short time	Does not exclude but correct the inconsistencies by adding "jnsq" variables Hardly achieve the goal to identify object based on knowledge of the object.	Involve complex method where for each inconsistent data, a "jnsq" variable will be added
Functional dependencies of corresponding relational variables	Manage to find inconsistent data in short time using functional dependencies between attribute's relation obtained from database	Step to correct inconsistent data are not included. Large dataset need more association rule and computing time	Bad database design can effect implementation of the technique User need to consider degree of dependency

		Not all database are designed with much consideration about normalization	to determine inconsistency
Fuzzy Multi-Attribute Decision Making	Can obtain high average correctness of data of data inconsistency solution	Do not discus on how to select key that will be used to identify similar objects Data source need to have good quality	Can only be detected while processing user queries Data quality criteria is important element in this technique
Fuzzy Multi-Attribute Decision Making	Has ideal performance compared to other services selection approach	Cannot easily obtained quality of service in open and future pervasive environment	Can only be detected while processing user queries Data quality criteria is important element in this technique

Based on Table 1 above, FMDAM is most suitable technique can implement in this research to overcome inconsistent data. The details implementation of FMDAM technique will describe in Section 3.

3. Implementation of Fuzzy Multi Attribute Decision Making

This section described about the overview of process involved in conducting in this research. Each process was explained in detail in development phase. Model architecture and system flow chart were illustrated in order to provide a clear view on components involved in this research together with the process flow. Technique selected as solution strategy was explained here briefly using the formula. Also in this chapter is system requirement that were used while completing this research.

3.1 Steps in Reducing Inconsistent Data

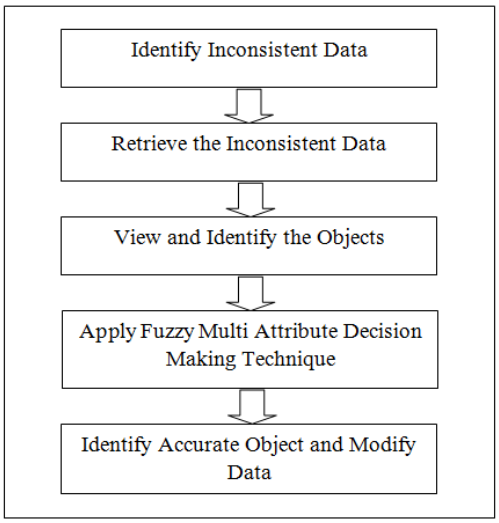

Figure 1. Steps in Reducing Inconsistent

Step 1: Identify Inconsistent

In this step, client's address was chosen as the target of the experiment. One state will be chose in order to identify the object of inconsistent data.

Step 2: Retrieve the Inconsistent Data

Data were retrieved by using 2 values as keyword which is the state and postcode the represents the data.

Step 3: View and Identify the Objects

In this step is to view the consistent data variance out of the data that has been retrieved from previous step. The objects that were to be used in the experiment will be selected.

Step 4: Apply Fuzzy Multi Attribute Decision Making Technique

Fuzzy multi attribute decision making technique was chosen to apply in this research. This technique used two quality vectors which are time and cost.

Step 5: Identify Accurate Object and Modify Data

In the last step is identifying accurate data object and modifying all inconsistent data based on the data object selection which was produced from the technique's implementation in previous steps.

3.2 Process flows for Reducing Inconsistent Data Using Fuzzy Multi Attribute Decision Making

Figure 2 shows the process flow for reducing inconsistent data using fuzzy multi attribute decision making. Three major processes in Figure 2 which are retrieve data, identify inconsistent data and modify data to standardized format before transfer all clean data into database.

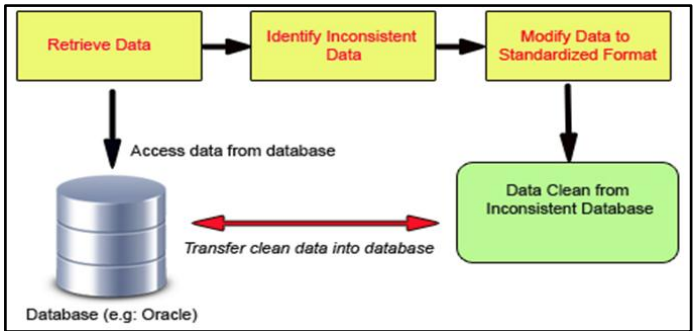

Figure 2. Process Flows

First process is retrieve data from database. This process will retrieve all data from database. After that, proceed with second process. In second process is quite difficult because to identify only inconsistent data. In order to identify inconsistent data, fuzzy multi attribute decision making technique was applied. In this technique, three important formula was implemented which are the quality matrix scale for positive and negative criteria formula, weighted Euclidean distance to positive and negative ideal solution formula and degree of

membership formula. Based on the result from this calculation, the best data object will select as a solution strategy for the inconsistency where later will be set as a standard format for all inconsistent data that represent the same attribute. Third process is modifying data to standardized format. In this process, some adjustments were made to the original data to ensure that all of the inconsistent data were stored in a new standardized format that similar to each other. The purpose of this process is to ensure that the inconsistency rate of the data will reduce. Finally, clean data will produce after implementation of these three important process and all the clean data will transfer into database. If inconsistency data still exist, the repetition of these processes must be proceed until all inconsistency data not exist anymore.

3.3 Fuzzy Multi Attribute Decision Making Algorithm

Fuzzy Multi Attribute Decision Making algorithm is based on Quality of Service (QoS). Two qualities that will be pointed out in this research are cost and time. The definition of cost is the amount of money that a service requester has to pay for executing the operation .As for time; it can be defined as the time taken for instance/objects while being processed. Time can also be identified as common measure for performance.

The decision making process consists of 5 steps:

Step 1: Derive quality vector for each object and obtain quality matrix where it rows represents the quality of service (QoS).

Step 2: Scale the quality matrix based on quality criteria. Since, this research focus on negative criteria which are cost and time therefore equation (1) will be used. For negative criteria, the higher the value, the lower quality it is.

Equation (1): $Vij = 1 - \left(\frac{qij}{\vee_i qij + \wedge_i qij}\right)$

From this step, a new set of quality matrix will be obtained where each row represents relative quality of objects.

Step 3: Compute weighted Euclidean Distance to positive and negative ideal solution.

Definition (1): Quality vector for positive ideal solution is the maximum number for each quality vector g = (g1,g2,g3,g4,g5) = $\vee_i^v i1, \vee_i^v i2, \vee_i^v i3, \vee_i^v i4, \vee_i^v i5$

Definition (2): Quality vector for negative ideal solution is the minimum number for each quality vector b = (b1,b2,b3,b4,5) = $\wedge_i^v i1\,, \wedge_i^v i2, \wedge_i^v i3, \wedge_i^v i4, \wedge_i^v i5$

Definition (3): Weighted Euclidean Distance between object and positive ideal solution

$$dig = \sqrt{\sum_{j=1}^{2} [wj(gj - vij)]^2}$$

Definition (4): Weighted Euclidean Distance between object and negative ideal solution

$$dib = \sqrt{\sum_{j=1}^{2} [wj(vij - bj)]^2}$$

Step 4: Calculate degree of membership of each object belonging to positive ideal solution
Degree of membership

$$Set\ vector, = \frac{1}{1+\left[\frac{dig}{dib}\right]^2}$$

Step 5: Select final decision alternative

The most minimum value from the result of calculation degree of membership is the most optimal answer. Once degree of membership for each object has been collected, it will be sort and the final decision would be the object with the minimum value of membership degree.

4. Implementation and Result

In this section, an application was developed using JAVA to test the capability and performance of Fuzzy Multi Attribute Decision Making. Oracle was used as a tool to manage and maintain the data in database.

4.1 Experiment and Result

In this section, experiment had been done to test capability of fuzzy multi attribute decision making algorithm in reducing inconsistent data. In these experiments, three functional processes involved which are:

i. Query the inconsistent data that under category of state using postcode and state as keywords.

ii. Each consistent data are defined with value of time criteria and cost criteria (quality matrix is obtained).

iii. Scale the quality matrix based on quality criteria using formula. Figure 3 show the code and example:

```
for(int i=0;i<time.size();i++){
        float a,a1,b;
                a=time.get(i);
                b=minTime+maxTime;
                a1=1-(a/(b));
                a1=round(a1,6);
                System.out.println(a1);
                time.set(i, a1);
}
```

Figure 3. Quality Matrix Code and Example

iv. Compute weighted Euclidean Distance to positive and negative ideal solution. Figure 4 shows the code and example:

```
for(int i=0;i<time.size();i++){
        float a,a1,b;
                a=time.get(i);
                b=minTime+maxTime;
                a1=1-(a/(b));
                a1=round(a1,6);
                System.out.println(a1);
                time.set(i, a1);
}
```

Figure 4. Weighted Euclidean Code and Example

v. Calculate degree of membership of each object belonging to positive ideal solution degree of membership.

vi. Select the final decision alternative where the most minimum value of degree of membership represents the most optimal object as a solution strategy.

vii. The inconsistent data will be modified to a standardized format based on the most optimal object that obtained from previous page.

Experiment

In this experiment, the inconsistent data that have been chosen as object for this research is state "Selangor". One column from the table, CLIENT_ADDRESS4 is selected as target data.

Table 2. Object of Inconsistent Data Selangor

S1	68100 BATU CAVES SELANGOR D.E.
S2	SELANGOR
S3	47100 PUCHONG,SELANGOR
S4	47100 PUCHONG
S5	47520 SELANGOR
S6	SELANGOR DARUL EHSAN
S7	43200 BATU 11, CHERAS
S8	SELANGOR D.E.

Step 1: Derive quality vector for each object and obtain quality matrix where it rows represents the quality of service (QoS).

Table 3. Object with Value of Quality Criteria

Quality Criteria	Cost	Time
S1	719	850
S2	719	21120
S3	719	2775
S4	719	25
S5	719	11.11
S6	719	15318.18
S7	719	11.11
S8	719	725

Step 2: Scale the quality matrix based on quality criteria.

Table 4. Object with Quality Criteria Value after Scaling Operation

Quality Criteria	Cost	Quality Criteria for Cost	Time	Quality Criteria for Time
S1	719	0.5	850	0.959775
S2	719	0.5	21120	0.000526
S3	719	0.5	2775	0.199789
S4	719	0.5	25	0.998817
S5	719	0.5	11.11	0.999474
S6	719	0.5	15318.18	0.275089
S7	719	0.5	11.11	0.999474
S8	719	0.5	725	0.96569

Calculation as below:

Quality criteria for time:
$1-\left(\frac{719}{719+719}\right)$ = o.5

Quality criteria for cost:
$1-\left(\frac{850}{21120+11.11}\right)$ = o.959775

Step 3: Compute weighted Euclidean Distance to positive and negative ideal solution.

Table 5. Positive and Negative Ideal Solution after Calculate Euclidean Distance

Quality Criteria	Quality Criteria for Cost	Quality Criteria for Time	Positive Ideal Solution	Negative Ideal Solution
S1	0.5	0.959775	0.007939	0.191849
S2	0.5	0.0.000526	0.199789	0
S3	0.5	0.868677	0.026519	0.17363
S4	0.5	0.998817	0.00001314	0.199658
S5	0.5	0.999474	0	0.199789
S6	0.5	0.275089	0.144877	0.054913
S7	0.5	0.999474	0	0.1999789
S8	0.5	0.96569	0.0065568	0.193233

Calculation as below:

Positive ideal solution:

$$\sqrt{[0.3(0.5-0.5)]^2+[0.2(0.999474-0.959775)]^2}=0.0079398$$

Negative ideal solution:

$$\sqrt{[0.3(0.5-0.5)]^2+[0.2(0.959775-0.000526)]^2}=0.0079398$$

Step 4: Calculate degree of membership of each object belonging to positive ideal solution

Table 6. Object with Membership of Degree Vector

Quality Criteria	Positive Ideal Solution	Negative Ideal Solution	Membership of Degree Vector
S1	0.0079398	0.191849	0.9982902
S2	**0.199786**	**0**	**0**
S3	0.026159	0.17363	0.9778049
S4	0.000314	0.199658	0.9999996
S5	0	0.199789	1.0
S6	0.144877	0.054913	0.125617
S7	0	0.199789	1.0
S8	0.006568	0.193233	0.998849

Calculation as below:

Membership of degree vector:

$$\frac{1}{1+(\frac{0.199786}{0})^2}=0$$

Step 5: Select final decision alternative

The minimum number in degree of membership represents the most optimal choice as a solution to inconsistent data. Since the minimum value is 0.0, S2 that represents the object "**SELANGOR"** is selected as the most optimal object.

4.2 Analyze the Implementation of Fuzzy Multi Attribute Decision Making

Based on experiments above, one column from table "CLIENT_ADDRESS4" was selected. In this column, name of objects are different, but referred to same place or location. In fuzzy multi attribute decision making have four steps to identify which is correct object. In table 3, two criteria was selected which are cost and time. The purpose of this selection is to calculate value of cost and time for each data by using specified formula in selected column. In table 3, cost value from S1 until S8 equal to 719. Meanwhile, time value for S5 and S7 is similar. The time value for S5 and S7 is 11.11. This is minimum value compared to others value. The cost and time value will use to calculate quality criteria for cost and quality criteria for time value by using specified formula. In Table 4, quality criteria for cost value started from S1 until S8 are equal to 0.5. Meanwhile, the quality criteria for time value are between 0 and 1 for S1 until S8. Negative value is last step to determine which one is correct object. Table 6 indicates the average negative ideal solution value for S1 until S8 except S2 are between 0 and 0.2. But the negative ideal solution value and membership of degree vector value for S2 is equal to 0. In this case, S2 is a correct object because of membership of degree vector value is minimum compared to others. Based on this analysis of the result above, fuzzy multi attribute decision making is suitable to reduce inconsistent data during accessing data from database.

5. Conclusion & Future Work

In conclusion, some of techniques in reducing inconsistent data from database have been studied. Fuzzy multi-attribute decision making was chosen to solve the problem statement had been stated in section 1. Implementation of this technique has four steps in order to determine which one is correct object. Result from experiments in section 4, indicates this technique is suitable to overcome inconsistent data during accessing data from database. In our future work, this technique will use to reduce inconsistent data from heterogeneous database access.

Acknowledgments

Special thanks to Universiti Sultan Zainal Abidin (UniSZA) and my friend Che Mat Ismail for his support and advice.

References

[1] M. K. Yusof and A. Azlan, "Comparative Study of Techniques in Reducing Inconsistent Data", International Journal of Database Theory and Application, vol. 5, no. 1, **(2012)** March.
[2] X. Wang, L. P. Huang, X. H. Yu and J. Q. Chen, "A Solution for Data Inconsistency in Data Integration", Journal of Information Science and Engineering, vol. 27, **(2011)**, pp. 681-695.
[3] Z. Pawlak, "Rough Set and Data Analysis", Proceedings of the Asian, **(1996)** December 11-14, pp. 1-6.
[4] Z. Pawlak, "Rough Set" International Journal of Computer and Information Science, vol. 11, no. 5, **(1982)**, pp. 341 – 356.
[5] H. Sug, "An Efficient Method of Data Inconsistency Check for Very Large Relations", International Journal of Computer Science and Network, vol. 7, no. 10, **(2007)**.

[6] H. Sug, "A Rough Set Based Data Inconsistency Checking Method for Relational Databases", International Journal of Computer Science and Network, vol. 8, no. 11, **(2008)** November.
[7] Z. Pawlak, "Rough Set", International Journal of Computer and Information Science, vol. 11, no. 5, **(1982)**, pp. 341 – 356.
[8] L. Cavigue, A. B. Mendes and M. Funk, "Logic Analysis of Inconsistent Data (LAID)", **(2010)**.
[9] H. Sug, "An Efficient Method of Data Inconsistency Check for Very Large Relations", International Journal of Computer Science and Network, vol. 7, no. 10, **(2007)**.

Authors

Mohd Kamir Yusof obtained her Master of Computer Science from Faculty of Computer Science and Information System, Universiti Teknologi Malaysia in 2008. Currently, he is a Lecturer at Department of Computer Science, Faculty of Infomatics, Universiti Sultan Zainal Abidin (UniSZA), Kuala Terengganu, Terengganu, Malaysia. His main research areas include information retrieval, database integration and web semantics.

Mohd Nordin Abdul Rahman obtained her PhD in Computer Science from Universiti Malaysia Terengganu in 2008. Currently, he is a Lecturer at Department of Computer Science, Faculty of Infomatics, Universiti Sultan Zainal Abidin, and Terengganu, Malaysia. His main research areas include web services, cloud computing and software engineering.

Atiqah Azlan obtained her Degree of Computer Science (Software Development) from Faculty of Informatics, Universiti Sultan Zainal Abidin in 2011. Currently, she is a Master Student at Department Computer Science, Faculty of Infomatics, Universiti Sultan Zainal Abidin, Terengganu, Malaysia.