Journal of Southeast University (English Edition) Vol. 20 No. 4 Dec. 2004 ISSN 1003—7985

Application of fuzzy equivalence theory in data cleaning

Li Huayang[1,2] Liu Yubao[1] Li Youkui[3]

([1] College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)
([2] UFsoft School of Software, Jiangxi University of Finance and Economics, Nanchang 330013, China)
([3] Nanjing Institute, Huawei Technologies Co., Ltd, Nanjing 210001, China)

Abstract: This paper presents a rule merging and simplifying method and an improved analysis deviation algorithm. The fuzzy equivalence theory avoids the rigid way (either this or that) of traditional equivalence theory. During a data cleaning process task, some rules exist such as "included"/"being included" relations with each other. The equivalence degree of the being-included rule is smaller than that of the including rule, so a rule merging and simplifying method is introduced to reduce the total computing time. And this kind of relation will affect the deviation of fuzzy equivalence degree. An improved analysis deviation algorithm that omits the influence of the included rules' equivalence degree is also presented. Normally the duplicate records are logged in a file, and users have to check and verify them one by one. It's time-cost. The proposed algorithm can save users' labor during duplicate records checking. Finally, an experiment is presented which demonstrates the possibility of the rule.

Key words: equivalence theory; equivalence degree; data cleaning

Data cleaning becomes an important task during the building process of a data warehouse. Nowadays many researchers have studied algorithms such as queue-prior algorithm and multi-pass sorted-neighborhood algorithm[1-6], which greatly increase the execution efficiency of data cleaning. However, the equivalence theory, a very important concept in data cleaning, is not given much attention. Equivalence theory is the theory of how to define two records to be equivalent or duplicated. Traditional equivalence theory claims that records that conform to given equivalent matching rules are duplicated; otherwise they are not. A new equivalence theory based on fuzzy theory[7] is introduced to improve data quality. This paper presents a rule-merging method to improve computing speed and puts forward an algorithm to cluster similar records. Finally, the paper gives an experiment and explains the experimental result.

1 Traditional Equivalence Theory and Fuzzy Equivalence Theory

The condition of a common set can be described with the characteristic function:

$$C_A(U) = \begin{cases} 1 & u \in A \\ 0 & u \notin A \end{cases}$$

The function shows the phenomena of "either this or that". The traditional equivalence theory looks like "that". A typical rule definition of traditional equivalence theory is as follows[4]:

Example 1

Given two records, R_1 and R_2,

If R_1.lname EQUALS, R_2.lname

And R_1.Addr EQUALS, R_2. Addr

And R_1.fname differs slightly R_1.fname

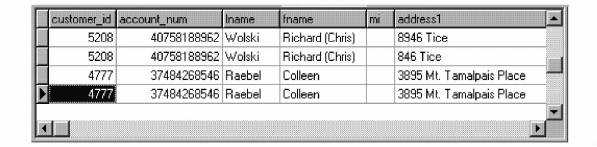
Then R_1 EQUALS R_2

In example 1, if the addresses and the first names of R_1 and R_2 are the same, and their second names are slightly different, then we regard R_1 and R_2 as duplicated records, otherwise they are not. That is, the traditional equivalence theory is in the form of "either this or that", if they are not duplicated records, they are non-duplicated.

Therefore, the traditional equivalence theory has two problems to be solved. Firstly, with such a mechanical rule, it is difficult to define whether two records are duplicated or not. Secondly, the quantity of rules in some data cleaning tools is too huge. Among so many rules, even though there exists no conflict, the choice of rules would be tedious. No matter how careful experts and users are, the generated rules can hardly handle well all the data problems in data warehouse. Therefore, some data-cleaning tools, given away to users to solve the data quality problems, demand interaction with users in the form of providing a LOG file to verify them (shown in Fig.1). But the LOG file would be so bulky that users may encounter data problems over time. So the data quality problem is unavoidable with the traditional equivalence theory.

Received 2004-01-09.

Biography: Li Huayang (1973—), male, doctor, associate professor, dariusli@tom.com.



customer_id	account_num	lname	fname	mi	address1
5208	40758188962	Wolski	Richard (Chris)		8946 Tice
5208	40758188962	Wolski	Richard (Chris)		846 Tice
4777	37484268546	Raebel	Colleen		3895 Mt. Tamalpais Place
4777	37484268546	Raebel	Colleen		3895 Mt. Tamalpais Place

Fig.1 LOG file in traditional data cleaning tools

In order to embody the continuous and the transitional features of such problems, ZadeH expanded the value range of the characteristic function from $\{0, 1\}$ to $[0, 1]$, **which is the fuzzy set.**

Definition 1 Fuzzy set

Assume a given mapping in domain U:

$$A: U \rightarrow [0, 1]$$
$$u \mapsto A(u)$$

Then A is the fuzzy set in U, $A(u)$ is the membership function of A.

We assign a value $[0, 1]$ for each rule, which we call equivalence degree. It means if two records conform to the rule, then they will be similar to each other with the probability of equivalence degree.

The user should define a threshold value through which users can determine which are the duplicated records among all records. And the union calculation method in fuzzy theory[7] is introduced to evaluate **the fuzzy equivalence degree.**

Definition 2 Assume A, $B \in F(U)$, then $A \cup B$ is the equivalence degree of A and B's, or the union of A and B's. Its membership function is

$$(A \cup B)(u) = A(u) \vee B(u) = \max(A(u), B(u))$$

After the new equivalence theory is adopted, many rules can be applied to records in the same group. So the rule conflicting problem is solved. Assuming there are three rules, each of which presents the corresponding duplicated records, and the possibility of concluding that R_1 and R_2 are matching is (0.9, 0.8, 0.85), respectively, and the threshold is 0.85. We get 0.9 with the union calculation method:

$$(\text{Rule1} \cup \text{Rule2} \cup \text{Rule3})(u) = \max(0.9, 0.85, 0.8) = 0.9$$

Therefore R_1 and R_2 are duplicated records. In fact, the fuzzy equivalence degree is determined by the maximum equivalence degree. If the equivalence value were defined to be larger, the duplication possibility of data conforming to the rule would be higher or vice versa. The feature must be reflected in the calculation **method of equivalence degree.**

2 Improvement on Fuzzy Equivalence Theory

2.1 Close degree and rule optimization

However, only when the union calculating method is insufficient, should an analysis deviation algorithm be introduced to analyze the deviation of fuzzy equivalence degree. Users may need some data to describe the influence of different rules on the same group. For example, if the equivalence degree of the three rules above on the other two records R_3, R_4 is (0.9, 0.2, 0.15), their fuzzy equivalence degree is also 0.9. Thus it cannot tell the difference between the first group data R_1 and R_2 and the second group data R_3 and R_4 with the equivalence degree. In fact, users tend to analyze the second group data, for their three data are greatly different.

A close degree function[7] is introduced to analyze the difference between the two groups. The equation is as follows (Hamming approximation accuracy):

If $U = \{u_1, u_2, \cdots, u_n\}$,

$$N(A, B) = 1 - \frac{1}{n}\sum |A(u_i) - B(u_i)|$$

Because the threshold is 0.85, let $B = (0.85, 0.85, 0.85)$, the close degree of the two groups of data is

$$N_1 = 1 - \frac{1}{3}(|0.85 - 0.9| + |0.85 - 0.8| + |0.85 - 0.9|) = 0.95$$

$$N_2 = 1 - \frac{1}{3}(|0.85 - 0.9| + |0.85 - 0.15| + |0.85 - 0.2|) = 0.53$$

From the above result, we can find that the approximation accuracy of the first group of data is larger, and of the second one is smaller. Thus the duplication possibility of the first group of data is higher, and the second group of data needs further observation.

During a data cleaning process, there are many rules to be used, among which there exist relations of "included" or "being included in". Assuming there are two rules, A and B, where the attributes included in rule B contain those included in rule A. It is obvious that users will give a higher equivalence degree to rule B than that to rule A. And we can draw a conclusion that the fuzzy equivalence of two records will be determined by rule B and the record-set from rule B is a subset of that from rule A.

With the threshold, there are two ways to optimize the data cleaning process.

1) Omit the rules of which the equivalence is lower than the threshold if users only care about the records with a higher probability being duplicated.

2) Check and group the rules, with "included" and "being included" relations. The program will process those rules together.

Users can choose the second way or both of them

to optimize the algorithm.

Another problem to be solved is the calculation of close degree if there are rules with the including relation. The close degree will be affected by two rules with the including relation. Omit the lower equivalence degree when calculating the close degree. For example, rule A and rule B make records R_1 and R_2 duplicated, where rule B includes rule A, we can omit the equivalence degree of rule A when we compute the close degree of the duplicated records between R_1 and R_2. The method will give us a close **degree with high quality.**

2.2 Duplicated records clustering algorithm

Another important problem is how we select correct data from duplicated records. For example, if the program detects that R_1 and R_2 are duplicated, how could it judge which one is accurate or both are problem data? A research tendency is to expect to realize automation to handle the problem. However, automation is difficult to realize because of different duplicated rewords. Instead, manual handling is still needed because obtaining data credibility through computer automation is an arduous process. It requires that enormous information be digitalized along with technical difficulties. And it demands a large quantity of interaction and information sharing between systems, which may cause conflict of interest problems.

Therefore manual interference is unavoidable. This paper addressed a clustering algorithm to handle duplicated records so as to reduce the quantity of manually handled information.

Our strategy is not to pick a record or insert it if it is a non-duplicated record one at a time. Instead, let users submit all non-similar records to the data warehouse after they browse and verify the records through duplicated records clustering. We assume R_1, R_2, R_3 are a group of duplicated records and R_4, R_5 are the other group among the five records R_1, R_2, R_3, R_4, and R_5. During the identification of duplicated records, the duplicated records are clustered and handled. Then R_1, R_2 and R_3 are clustered as a group and R_4 and R_5 are clustered as another group. When users browse and prepare to handle the duplicated record R_1, R_2 or R_3, the three records will appear simultaneously. Then users can find out their duplicated parts and causes of inconsistency. After being hand revised, the content that users choose will be inserted into the data warehouse and the other two records will be deleted from the original data set together with it. In this way it is unnecessary for users to browse R_1, R_2, R_3 in the sequence of LOG file and to worry about whether R_2 and R_3 have ever been handled. Therefore it is helpful for users to speed up **execution and ensure data quality.**

Clustering algorithm

Build an undirected graph G,

1) Each duplicated record is a node in the graph G.

2) If two records exist with similar relations, then link them with a line. If they have a link, skip it.

3) Repeat step 2) till all records are done.

Thus, we get a no-oriented graph. Each sub-graph is a cluster of duplicated records.

For example, if $R_1 \rightarrow R_2$, $R_1 \rightarrow R_3$, $R_2 \rightarrow R_5$, $R_6 \rightarrow R_4$, $R_7 \rightarrow R_8$, according to the clustering algorithm, we have an undirected graph G_1, as shown in Fig.2.

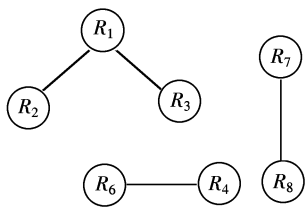


Fig.2 Undirected graph G_1 on duplicated records

From Fig.2, we can get three clusters from the duplicated records. The users may browse and deal with the duplicated records easily by a clustering algorithm. As shown in Fig.3, when we choose the 91st duplicated record, all the other duplicated records with the 91st record are presented in the following table. In Fig.3, duplicated records are presented. One's address1 value is "413 Miller Dr." and "41 Miller Dr." Actually they are inconsistent.

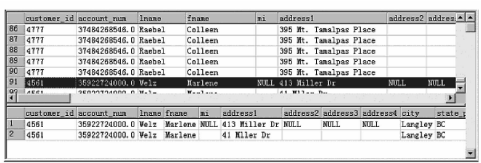


Fig.3 Result of clustering method

3 Experiment

In order to show the use of the equivalence theory, we develop a program and adopt the database named pubs attached in Microsoft SQL Server 2000. We choose 200 records in table customer among those records, some are the same in customer_name (fname), address, or telephone. Here we add a column called age and generate 50 duplicated records with a little different.

As mentioned above, we have to determine the equivalence degree according to the user's experience. Here we put forward four rules as follows:

1) The equivalence degree is 0.5 if the customer _ name is the same.

2) The equivalence degree is 0.85 if the customer _ name and address are the same.

3) The equivalence degree is 0.96 if the customer _ name, the address and the telephone are the same.

4) The equivalence degree is 0.86 if the age and address and customer _ name are the same.

Then we get the result shown in Tab.1.

Tab.1 Number of duplicated records corresponding to rules

Total duplicated records	Rule 1	Rule 2	Rule 3	Rule 4
50	65	53	51	54

As shown in Tab.1, most duplicated records are founded, where Rule1 ⊂ Rule2 ⊂ Rule3. Thus, the number of duplicated records founded by rule 1 will be larger than that founded by rule 2. Analogically the number of duplicated records founded with rule 2 is larger than that with Rule 3, …, which is supported by the experimental result. The computing time can be reduced by merging rule 1, rule 2 and rule 3. However, the number of founded duplicated records is not equal to the corresponding equivalence degree. How to determine the equivalence degree is for future **work.**

4 Conclusion

The traditional equivalence theory is in the form of "either this or that", which cannot reflect the fuzzy problems in reality and sets obstacles for users to define and generate rules. This paper introduces and improves the equivalence theory based on fuzzy theory, which not only adapts to fuzzy phenomena in reality, but also accords to semantic cleaning tendency to a certain extent. This paper puts forward an improved analysis deviation solution in fuzzy equivalence theory and presents an experiment. Moreover, this paper presents the clustering method of handling duplicated records. It helps users to complete **the clustering of duplicated data on time.**

References

[1] Rahm E, Hai Do H. Data cleaning: problems and current approaches[J]. *Data Engineering*, **2000, 23**(4): 3 - 13.

[2] Davidson Susan B, Kosky Anthony S. Specifying database transformations in WOL [J]. *Data Engineering*, **1999, 22** (1): 25 - 31.

[3] Haas Laura, Miller Renee, Niswonger Bartholomew, et al. Transforming heterogeneous data with database middleware: beyond integration [J]. *Data Engineering*, **1999, 22**(1): 31 - 37.

[4] Raman V, Joseph M. Potter's wheel: an interactive data cleaning system[A]. In: *Very Large Data Bases* [C]. ACM Press, 2001. 381 - 390.

[5] Galhardas Helena, Florescu Daniela, Shasha Dennis. Declarative data cleaning: language, model and algorithms [A]. In: *Very Large Data Bases* [C]. ACM Press, 2001. 371 - 380.

[6] Hernandez Mauricio A, Stolfo Salvatore J. The merge/ purge problem for large databases [A]. In: *SIGMOD Conf* [C]. ACM Press, 1995.127 - 138.

[7] Li Huayang, Liu Yubao, Li Youkui. The equivalence theory based on fuzzy theory [A]. In: *The* 3*rd International Conf on Machine Learning and Cybernetics* [C]. Shanghai, 2004. 1272 - 1276.

模糊等值理论在数据清理中的应用

李华昉[1,2] 刘玉葆[1] 李又奎[3]

([1] 华中科技大学计算机科学与技术学院, 武汉 430074)

([2] 江西财经大学用友软件学院, 南昌 330013)

([3] 华为技术有限公司南京研究所, 南京 210001)

摘要: 提出了规则合并的优化方法和重复记录聚类清除的方法. 应用模糊等值理论,避免了传统等值理论非此即彼的僵硬方式,但清理过程中部分规则可能存在包含与被包含的关系,被包含的规则其等值度显然会相对较小,根据用户阀值提出了规则合并的优化方法,可减少重复记录的计算时间. 基于同样的原因,规则间的包含与被包含关系将影响模糊等值度的误差分析,因此提出了利用忽略被包含的规则等值度提高误差分析精度的改进模糊等值理论误差分析方法. 重复记录的核实通常需要人工逐条检测,易于出错,本文提出的聚类算法,可节省大量的用户劳动. 最后给出一个实验,表明了规则优化的可能性.

关键词: 等值理论; 等值度; 数据清理

中图分类号: T391