International Journal of Computer Theory and Engineering, Vol. 2, No. 1 February, 2010
1793-8201

Data Quality Measurement using Data Mining

Saeed Farzi, Ahmad Baraani Dastjerdi

***Abstract*— Regarding to use of correct information in many applications, data quality measurement is very important. Nowadays, many statistical methods for measuring the quality of data have been proposed. The major problem of statistical methods is to lack of using data nature to measure the quality. Data mining is another method for measuring the quality of data. Data mining algorithms extract some knowledge. The extracted knowledge is used to measure the quality of data. In this paper, we introduce a new method, which uses data mining to extract some knowledge from database, and then we use it to measure the quality of input transaction.**

***Index Terms*—data quality, data mining, input transactions, statistical method.**

I. INTRODUCTION

The major role of most information systems is to present the real world in the computer world. People in the organization can create products or make decisions with information that is included in information systems. If this information does not compatible with the real world, systems will be poor and the organization, which uses them, will be began to act irrationality [2],[14],[15], [18]. Therefore, Data quality can be defined as consisting between data in the information system and that same data in the real world [2],[13],[14].

Data mining and statistical methods have been used to measure data quality. Statistical methods introduced some metrics, which they have been calculated by statistical functions such as average [2]. Mr. Ulrich G üntze and et al [3] have been used data mining algorithms to measure data quality. In their method, There are three steps for measuring data quality.1) Extract all association rules. 2) Select compatible association rules. 3) Add confidence factor of compatible rules as criteria of data quality of transaction. There are two important challenging issues. First, extracting all association rules needs a lot of time and next, there is no exact mathematical formula for measuring data quality.

In this paper, we introduce a method for measuring data quality based on data mining algorithms, which solves above challenging issues. Our method does not need to extract all association rules and we propose exact mathematical formula for measuring data quality.

In our method, we propose an algorithm with three steps, which calculates the data quality of transaction. step1: Extract association rules, which depend on input transaction (T) and are adapted by the functional dependency. Step2: Separate compatible and incompatible association rules. Step3: Calculate the quality of input transaction.

This paper is organized as follows. Notations and Definitions are explained in Sec.II. Data quality mining is described in Sec.III. Mathematical formula for measuring data quality is introduced in Sec.IV. Experimental study is covered in Sec.V. Concluding remarks are included in Sec.VI.

Manuscript received October 9, 2008.

Saeed Farzi is with Department of Computer Engineering, Islamic Azad University – branch of Kermanshah, Iran (corresponding author to provide phone: +988318247902; fax: +988317243065; e-mail: saeedfarzi@gmail.com).

Ahmad Baraani Dastjerdi is with the Computer Engineering Department, University of Isfahan, Iran. (e-mail: ahmadb@eng.ui.ac.ir).

II. NOTATIONS AND DEFINITIONS

D*: Database*

T: *Transaction that is included by the set of items*

I: *set of all items in D*

MIN_SUPP: *minimum support for set of items[4].*

Definition 1(association rule): *Let I is set of all items in D and X,Y are subset of I and* $X \cap Y = 0, Y \subseteq I, X \subseteq I$ *, so* $X \rightarrow Y$ *is called an association rule that Y is head and X is body of the association rule [4].*

Definition 2(support factor of X): *If X is a subset of I, the support factor of X (SUPP(X)) will be as follow [4]:*

$$\sup p(X) = \frac{|\{T \in D | X \subseteq T\}|}{|D|} \quad (1)$$

Definition 3 (support factor of an association rule)*: If X and Y are subsets of I and* $X \rightarrow Y$ *is an association rule, The support factor of* $X \rightarrow Y$ *(SUPP(*$X \rightarrow Y$*)) will be as follow [4]:*

$$\mathrm{SUPP}(X \rightarrow Y) = \mathrm{SUPP}(X \cup Y) \quad (2)$$

Definition 4(confidence factor of an association rule): *If X and Y are subsets of I and* $X \rightarrow Y$ *is an association rule, the confidence factor of* $X \rightarrow Y$ *(CF(*$X \rightarrow Y$*)) will be as follow [4]:*

$$\mathrm{CF}(X \rightarrow Y) = \mathrm{SUPP}(X \rightarrow Y)/\mathrm{SUPP}(X) \quad (3)$$

Definition 5(The association rule is adapted by functional dependency): *Let R:* $X \rightarrow Y$ *be an association*

rule and FD:A→B be a functional dependency where A and B are sets of database attributes, R will be adapted by FD if attributes in X are member of A and attributes in Y are member of B and vice versa.

Definition 6(association rules depend on a transaction (T)): *If T is a transaction and R: X→Y is an association rule, R is depended on T if* $X \subseteq T$.

Definition 7(compatibility): *If T is a transaction and R: X→Y is an association rule and* $X \subseteq T \Rightarrow Y \subseteq T$ *, R will be compatible with T.*

Definition 8(incompatibility): *If T is a transaction and R: X→Y is an association rule and* $X \subseteq T \Rightarrow Y \not\subset T$ *, R will be incompatible with T.*

III. DATA QUALITY MINING

The quality of database will be set in a desirable extent in long term by considering input transactions in databases and avoiding not to inter unqualified data. For achieving this goal, we should evaluate the quality of every transaction that inters some data to the database and rollbacks transactions, which have not good quality. In this way, we can measure data quality of input transaction with Algo.1

Algorithm 1 (Data Quality Measurement)
Step1: *Extract association rule, which depends on the input transaction (T) and is adapted by the functional dependency.* ***Step2:*** *Separate compatible and incompatible rules.* ***Step3:*** *Calculate the quality of input transaction.*

Step1*: Extract association rules, which depend on the input transaction (T) and are adapted by the functional dependency.*

Unlike Mr. Ulrich G ünztze and et al [3], in Algo.1 it is not necessary to extract all association rules of database. You must obtain large itemsets from database and use them to make association rules. Extracting all association rules needs a lot of time and too much memory (time and space complexity). If we reduce the number of large itemsets, we can decrease a number of association rules. Therefore, we decrease time and space complexity. In addition, in Algo.1, we have just extracted association rules, which depend on input transaction and are adapted by one of the functional dependency. Therefore, we will reduce the number of large itemsets.

Algorithm 2(extracting association rules which depend on input transaction (T))
Step1*: Extract power set of items in T* ***Step2:*** *Extract the large itemsets (condition for each itemset: one of power set member must be subset of the large itemset)* ***Step 3:*** *Extract association rules from the large itemsets [definitions 5, 6, 7] (condition: one of the functional dependencies must adapt Association rules.*

In Algo.2, the number of large itemsets will be reduced because of a power set member must be subset of the large itemsets. Also In third step, there is a limitation for extracting association rules from the large itemsets. It causes to extract fewer association rules.

Step 2: *Separate compatible and incompatible rules.*

In this step, the compatible and incompatible association rules must be distinguished by Algo.3.

Algorithm3(separate the compatible and incompatible association rules)
Repeat for all extracted association rules. - *If AR: body→head is an association rule and* $head \subseteq powerset(T)$ [1] *then AR will be compatible rule* - *If AR: body→head is an association rule and* $head \not\subset powerset(T)$ *then AR will be incompatible rule*

Step3: *Calculate the quality of input transaction*

In this step, we use Eq.1 to calculate the quality of input transaction.

IV. QUALITY MEASUREMENT FUNCTION

All extracted association rule (Algo.2) have been categorized into the compatible and incompatible rules (Algo.3).

Axiom: if n equals the number of association rules depend on input transaction (T) that is extracted from D and they are adapted with functional dependency in D. Eq1 that is given as follows will calculate the quality of T.

$$Q(T) = \frac{n - nc + \sum_{i=1}^{nc} cf_i - \sum_{j=1}^{n-nc} cf_j'}{n} \quad (4)$$

Where nc is the number of compatible rules. cf_i is the confident factor of the i_{th} compatible association rule and cf_j' is the confident factor of the j_{th} incompatible association

rules.

Proof:

Let r: body→head be an association rule with cf as a confident factor. Based on definition 2, 3, 4, we can calculate $cf(r)$ as follow.

$$cf(r)=\frac{\sup(body\rightarrow head)}{\sup(body)}=\frac{\sup(body\cup head)}{\sup(body)}=$$

$$\left(\frac{\left|\{T\in D\,|\,(body\cup head)\subseteq T\}\right|}{|D|}\right)/\frac{\left|\{T\in D\,|\,body\subseteq T\}\right|}{|D|}$$

$$cf(r)=\frac{\left|\{T\in D\,|\,(body\cup head)\subseteq T\}\right|}{\left|\{T\in D\,|\,body\subseteq T\}\right|} \quad (5)$$

Eq.2 shows that, $cf(r)$ presents probability of correctness of *r*. Therefore, compatible association rules have a positive effect on transaction (T) so we must add their cf s.

$$Q(T)=cf_1+cf_2+\ldots+cf_{nc}=\sum_{i=1}^{nc}cfi \quad (6)$$

In addition, incompatible association rules have a negative effect on T. let r'=body →head with cf' be incompatible rule. $cf'(r')$ presents probability of correctness of r'. Therefore, $1-cf'(r')$ presents probability of incorrectness of r' that has a positive effect on correctness of transaction (T).

$$Q(T)=\sum_{i=1}^{nc}cf_i+\sum_{j=1}^{n-nc}(1-cf_j') \quad (7)$$

By dividing the Eq.7 to the number of all extracted association rules (n). So Eq.8 is concluded.

$$Q(T)=\frac{\sum_{i=1}^{nc}cf_i+\sum_{j=1}^{n-nc}(1-cf_j')}{n} \quad (8)$$

After summarizing Eq.8, Eq.4 is achieved.

V. Experimental Study

For describing proposed method, we illustrate a case study. There is a table that is called Employee (state, zip code, city, job, salary) and there are functional dependencies that are shown as follows.

State →Zipcode

City →State

Job →Salary

If *"State=Kermanshah, Zipcode=831, city= Biseton, Job=manager, Salary=1000"* is input transaction (T), the following association rules depend on it will be extracted they are shown as follows.

State=Kermanshah →Zipcode=831 with cf=1.0

City=Biseton →State=Kermanshah with cf= 1.0

Job=manager →Salary=1000 with cf= 1.0

Some of these association rules are compatible and some of them are incompatible with T. Compatible association rules with T are shown as follows (Algo.3) (Definition: 7):

State= Kermanshah →Zipcode=831 with cf=1.0

City= Biseton →State= Kermanshah with cf=1.0

Incompatible association rules with T are shown as follow (Algo.3)(Definition 8):

Job=manager →Salary=1000 with cf=1.0

Therefore, By using Eq.1, the quality of the transaction (T) equals 66%. The quality of one hundred transactions has been calculated by proposed method (Algo.1) and has been compared with results from human expert. The results are shown in Fig.1.

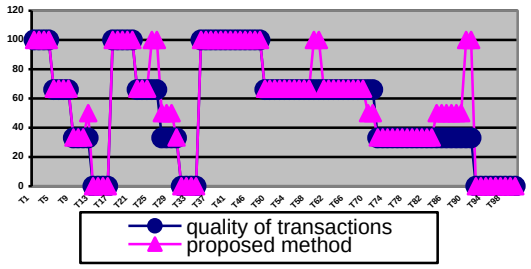

Fig.1 The quality of one hundred transactions has been calculated by proposed method (Algo.1) and has been compared with results from human expert

The proposed method calculates the quality of 84 cases correctly (84%). In order to evaluate the proposed method more precisely, we use 10000 transactions. The quality of 7135 cases is calculated correctly (71%). In addition, these transactions have been calculated by Ulrich G ünтze and et al [3] method that 50% of them have been calculated correctly.

VI. Conclusion

In addition to statistical methods, data mining methods can measure the quality of the data. This method is called data quality mining that uses the nature of the data in measuring the quality. DQM can be a suitable privilege compared to the statistical methods. In this paper, we introduce a new method, which uses data mining to measure the quality of transaction. The quality of transaction is calculated by Eq.1. Our method is successfully applied to case study in measuring quality of transaction. Measuring data quality of transaction obtained from our method showed a good general agreement with the results from human expert.

References

[1] Strong, D.M,Lee Y.W,wang, R.Y., "Data Quality in Contex", communication of ACM, 40(5),1997.

[2] Pipino,L.,lee,W., Wang,y., "Data Quality Assessment"1998.

[3] Hipp ,Jochen, G ünтzer, Ulrich, Grimmer, Udo, "Data Quality Mining",2002.

[1] Poweset is set of all subset of T

[4] Agrawal,R.,Srikant,R. "Fast Algorithm for Ming Association Rules",In proc of the 20th Int'l conference on very larg database, santiago ,1994.

[5] Agrawal,R., Imielinski,T., Swami,A., " Mining Association Rules between Sets of Items in Large Databases", In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., 1993.

[6] Anwar,T., Navathe,T., Beck,T., "Knowledge Mining in Databases: a Unified Approach through Conceptual Clustering." , Georgia Institute of Technology, 1992.

[7] Agrawal,R., Srikant,I., "Mining Sequential Patterns.",In *Proc. of the Int'l Conf. on Data Engineering (ICDE)*,Taipei, Taiwan, 1995.

[8] Sabrina Vazquez Soler, Daniel Yankelevich: Quality Mining: A Data Mining Based Method for Data Quality Evaluation. In *MIT Conference on Information Quality (IQ)*, 2001,pp.162-172.

[9] Yair Wand, Richard Y. Wang: Anchoring Data Quality Dimensions in Ontological Foundations. CACM ,39(11),1996,pp. 86-95.

[10] Richard Y. Wang, Henry B. Kon, Stuart E. Madnick: Data Quality Requirements Analysis and Modeling. In, *Proceedings of the Ninth International Conference on Data Engineering*, 1993,pp. 670-677.

[11] Richard Y. Wang: A Product Perspective on Total Data Quality Management. CACM 41(2), 1998, pp. 58-65.

[12] Richard Y. Wang, Diane M. Strong: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4), 1996,pp. 5–34.

[13] Richard Y. Wang, Veda C. Storey, Christopher P. Firth: A Framework for Analysis of Data Quality Research. IEEE Transactions on Knowledge and Data Engineering 7(4),1995,pp. 623-640.

[14] Richard Y. Wang, Mostapha Ziad, Yang W. Lee: Data Quality. Kluwer 2001.

[15] Jack E. Olson: Data Quality: The Accuracy Dimension. Morgan Kaufmann ,2003.

[16] Barbara Pernici, Monica Scannapieco: Data Quality inWeb Information Systems. In *ER 2002, 21st International Conference on Conceptual Modeling*, LNCS 2503, 2002, pp. 397-413.

[17] Markus Helfert, Eitel von Maur: A Strategy for Managing Data Quality in Data Warehouse Systems. In *MIT Conference on Information Quality (IQ)*, 2001,pp. 62-76.

[18] Theodore Johnson, Tamraparni Dasu: Data Quality and Data Cleaning: An Overview. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 681, 2003.

Saeed Farzi was born in Kermanshah in 1983. He is an faculty member at the Department of Computer Engineering, Islamic Azad University-beranch of Kermanshah, Iran. He received his B.S. in Computer Engineering from Razi university in Iran , in 2004 and M.S. in Artificial Intelligence from Isfehan university in Iran, in 2006. His current research interests include artificial intelligence , Neural Network , soft computing and Grid computing..