RESEARCH ARTICLE OPEN ACCESS

Synthesizing High-Frequency Rules from Various Data Sources Using Weighting Method (Determining Service Quality in Hospitals)

Mr. A. Arulmurugan [1], A. Ajitha [2], R. Gracy [3]
M.Tech. Associate Professor [1]
Department of Information Technology
A.V.C. College of Engineering, Mayiladuthurai
Tamil Nadu - India

ABSTRACT
Many large organizations have multiple data sources, while putting all data together from different sources might a mass a huge database for centralized processing. A weighting method is proposed in this paper for identifying valid rules among the large number of forwarded rules from different data sources. Valid rules are the rules which are supported by most of the branches of an organization. A hospital may consist of number of branches which are located in different places. Every branch having different services. It is highly difficult to assess the services offered by various branches in manual manner and hard to identify the strength and weakness of every branch. To avoid such drawback the proposed project aims to determine service quality in all branches using data mining techniques. Hence this method is applied to rank the rules based on patient perceived service qualities in a hospital. Experimental results show that this proposed weighting model is efficient and effective.
Keywords: — Association based data mining, Data reduction, weights, and SERVQUAL scale.

I. INTRODUCTION

The aim of this paper is to design an algorithm for extracting valid rules among the large number of forwarded rules from different data sources by a weighting method. Data Mining is as an analysis of information that can extract useful patterns from large databases, has been widely applied to analyze data for decision makers. Data Mining, a1so known as knowledge discovery in databases, aims at the discovery of useful information from large collections of data. The discovered knowledge can be referred to as rules describing properties of the data, frequently occurring patterns, clustering of objects in the data base which can be used to support various intelligent activities such as decision making, planning and problem solving. Let I ={ i1,i2 .i3,….in }be a set of N distinct literals called items, and D be a set of transactions over I. Each transaction contains a set of items i1,i2, i3,….ik€ I. A transaction has an associated unique identifier called TID (Transaction Identification Number). An association rule is an implication of the form A→B, where A, B I, and A∩B =null set. A is called the antecedent of the rule, and B is called the consequent. A set of items (such as the antecedent or the consequent of a rule) is called an itemset. Each itemset has an associated rule Synthesizing rules is the process of putting all rules together and to produce valid rules from that. To mine transaction databases for large organizations that have multiple data sources, there are two possible ways.

(i) putting all data together from different sources to amass a centralized database for centralized processing, possibly using parallel and distributed mining techniques.

(ii) reusing all promising rules discovered from different data sources to form a large set of rules and then

searching for valid rules that are useful at the organization level. There are many methods and algorithms suggested for this second task. FP-tree-based frequent patterns mining method was developed by Han. This method is found efficient than the Apriori algorithm also an OPUS - based algorithm has been reported by Webb to reduce the searched space by focusing association rules mining with which the searched space consists of all possible items and item sets in a database. The Apriori algorithm uses a two step technique to identify association rules, and a search space in Apriori consists of all items and possible itemsets. However, existing work has focused on mining frequent itemsets in data sets, and few research efforts have been reported on post mining that gathers, analyzes, and synthesizes association rules from different data sources. Xindong tried a technique for this synthesize problem and came out with a solution of normalizing the weights of data sources proposed by Xindong.

II. PROBLEM SOLVING

The patients in the branches (6 in total) of a reputed private hospital in North India were questioned on the service qualities of the hospital using 12 item service quality scale, adapted from the SERVQUAL scale. Total size of the data source was 100 records. A survey questionnaire is constructed incorporating 12 service quality items (as presented in Table1) for assessing the influence of each indicator on the performance of Hospital services. The questionnaire has been

made simple and easy to understand. It is a closed-ended questionnaire based on Servqual scale. Questionnaire is enclosed in Annexure-I. The result of mining the above datasets using the factors such as patient oriented, competence, tangibles and convenience, is given below as four different association rules.

R1 :- Patient oriented

The rule patient oriented comprised of variables like extent of prompt service, consistent courtesy and knowledge to answer patient's questions, t, operating hours, grievance handling time, privacy. Then the rule is expressed as follows:

If V2 = yes AND V3 = yes AND V4 = yes AND V12 = yes then patient oriented = yes.

R2 :- Technical Competence

Variables like extent of interest in problem solving, right service, infection control, facilities available combined to define "technical competence". Then the rule is expressed as follows:

If V1 = yes AND V5 = yes AND V7 = yes AND V9 = yes then Technical Competence = yes.

R3 :- Empathy

The empathy factor consisted of variables like the extent of the firm's individualized attention to its patients. Then the rule is expressed as follows:

If V8 = yes AND V10 = yes the Empathy = yes.

R4 :- Convenience

Variables like guidance signs and timeliness and subsidized cost of medicine combined to define "convenience". Then the rule is expressed as follows:

If V6= yes AND V11 = yes then Convenience = yes.

The Table gives the local support and local confidence values for the rules Rl, R2, R3 and R4. Local support and local confidence define the support and confidence of the local rules respectively. The problem here is when these association rules are forwarded from different known data sources in the branches of a hospital to their headquarters, it requires a method to synthesize these association rules for knowing the valid rules among them. The Fig. 1 illustrates this model.

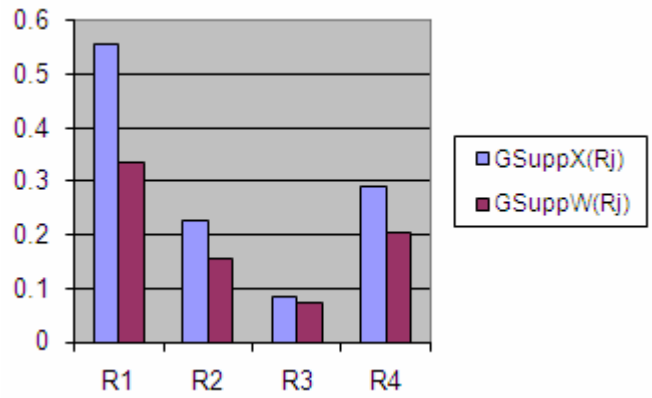

Fig. 1. Performance Analysis Bar Diagram

Let D1,. D2 ,... , Dm be m different data sources from the branches of a large hospital (name could not disclosed due to confidential reasons) of similar size, and Si be the set of association rules from Di (i = 1,2, ... , m). Also, let W1 W2, W3, ... , Wm, be the weights of these data sources. Then for a given rule Ri, expressed as X Y, its global support and global confidence are defined as follows:

$$Gsupp(Ri): \sum_{i=1}^{m} wDi * Lsupp(Ri)$$

$$Gconf(Ri): \sum_{i=1}^{m} wDi * Lconfi(Ri)$$

where Gsupp(Ri) is the support of Ri after synthesizing, Gconf(Ri) is the confidence of Ri after synthesizing, Lsuppi(Ri) is the support of Ri in Di and Lconfi(Ri) is the confidence of Ri in Di, i = 1, 2, ... , m. The synthesis of rules in our model is generally straightforward once all weights are reasonably assigned. The weight of each rule is calculated by its frequency in the original data sources.

III . ARCHITECTURE DIAGRAM

IV. ALGORITHM

This section first discusses on the informal description of the synthesizing rules by weighting algorithm developed for ranking the rules based on patient perceived service qualities in a hospital. Next, it presents the formal description of the algorithm. Finally, it provides the impact of the algorithm.

4.1 Description

In order to synthesize association rules from different data sources in the branches of a hospital, this method needs to determine the weight for each data source. In our opinion, if all data sources are of similar size, the weight of each data source can be determined by the rules discovered from it. Here the data sources are the six branches of a reputed hospital in North India.

Let D1,. D2 ... , Dm, be m different data sources in the branches of a hospital, Si the set of association rules from Di(i= 1, 2, ... , m), and S = {S1,S2,S3,S4 .... Sm}.This method takes the frequency of a rule Ri in S to assign a rule weight wRi. The inter support relation between a data source and its rules can be applied to assign the data sources a weight. If a data source supports a larger number of high-frequency rules, the weight of the data source should also be higher. Table2 illustrates the above idea with the data. Let minsupp= 0.25.

minconf = 0.42, and the following rules be mined from three different branches of the hospital.

For the illustration purpose, we assume only three branches.

Datasource Dl :- S1 = {R1, R2, R4}

R1 with Lsupp = 0.50, Lconf = 0.27

R2 with Lsupp = 0.31, Lconf = 0.30

R4 with Lsupp = 0.47, Lconf = 0.82

Datasource D2:-

S2 = {R1, R4, R3}

R1 with Lsupp = 0.40, Lconf= 0.69 ;

R3 with Lsupp = 0.30, Lconf = 0.60 ;

R4 with Lsupp = 0.27, Lconf = 0.59 ;

Data source D3:-

S3 = {R2, R1}R1 with Lsupp = 0.43, Lconf = 0.73 ;

R2 with Lsupp = 0.31, Lconf = 0.71;

Thus S = {S1, S2, S3}. Here, the number of sources that contain Rl = 3, R2 = 2, R3= 1, and R4=2.

We can use the frequency of a rule in S to assign a weight for rules. The weights are assigned as follows:

wR1 = 3/ (3+2+1+2) = 3/8 = 0.375 ;

wR2= 2/(3+2+1+2) = 2/8 = 0..25 ;

wR3= 1/(3+2+1+2) = 1/8 = .125;

wR4= 2/(3+2+1+2) = 2/8 = 0.25 ;

We have seen that rule R, has the highest frequency and it has the highest weight; rule R3 has the lowest frequency and it has the lowest weight. Let S ={S1,S2,S3, ... ,Sm }, and R1, R2, ... Rn be all rules in S. Then, the weight of a rule Ri is defined as follows:

$$wRi = \frac{Num(Ri)}{\sum_{j=1}^{n} Num(Rj)}$$

where i = 1,2, ... , n ; and Num(R) is the number of data sources that contain rule R, or the frequency of R in S.

If a data source has a larger number of high-frequency rules. the weight of the data source should also be higher. If the rules from a data source are rarely present in other data sources, the data source would be assigned a lower weight.

To implement this argument, we can use the sum of the rule's weights divided by total number of data sources.

wD1 = (.375 + 0.25 +0.25)/3 = 0.2917

wD2 = (.375 + 0.125 + 0.25)/3 = 0.25 ;

wD3 = (0.375 +0.25)/3 = 0.2083 ;

Let D1,D2,D3, ... ,Dm be m different data sources in the branches of a hospital, S the set of association rules from D.(i = 1,2, ... , m), S ={S1, S2, ... , Sm }and R1,R2,... Rn be all rules in S. Then, the weight is defined as follows:

$$wDi = \frac{\sum_{i=1}^{n} wRi}{m}$$

After all data sources have been assigned weights, it requires a synthesize process to evaluate these association rules. Hence this paper introduces a simplified formula for computing global support and global confidence to replace the normalization process formula proposed by Xindong.

For Rule R1 : Patient oriented

Gsupp(R1) = wD1 * Lsupp1(R1) + wD2 Lsupp2(Rl) + wD3 * Lsupp3(Rl)

=0.2917* 0.50+ 0.25 * 0.40+ 0.2083 *0.43

=0.3354

Gconf(R1) =wD1 * Lconf1(R1) + wD2 * Lconf2(Rl) + wD3 * Lconf3(Rl)

=0.2917* 0.27+ 0.25*0.69 + 0.2083 * 0.73 =0.4033

For Rule R2: Technical Competence

Gsupp(R2)=wD1*Lsupp1(R2)+ wD3*Lsupp3(R2)

=0.2917* 0.31+ 0.2083 * 0.31

=0.155

Gconf(R2)= wD1 * Lconf1(R2) + wD3 * Lconf3(R2)

=0.2917* 0.30 + 0.2083 * 0.71

=0.2354

For Rule R3 : Empathy

Gsupp(R3) == wD2 * Lsupp2(R3)

= 0.25 * 0.30

= 0.075

Gconf(R3)= wD2 * Lconf2(R3)

= 0.25 * 0.60

= 0.15

For rule R4 : Convenience

Gsupp**(**R4**)=** wD1 * Lsupp1(R4) + wD2 * Lsupp2(R4)

= 0.2917* 0.47+ 0.25 * 0.27

= 0.2045

Gconf**(**R4**)**= wD1* Lconf1(R4) + wD2 * Lconf2(R4)

=0.2917* 0.82+ 0.25 *

0.59

= 0.3866

Thus the ranking of the rules by their global supports is R1, R4, R2 and R3. According to this ranking, we can select high-

rank rules after the minimum support and minimum confidence. Table 3 gives the calculated value of Gsupport and Gconfidence for the rules using Xindong method. Table 4 gives the calculated value of Gsupport and Gconfidence for the same rules using the proposed method. This procedure is transformed into an algorithm in the next section.

4.2 Formal Description

Algorithm :- Synthesizing Rules By Weighting method

Input:

S= {S1,S2,....Sm} : rule sets ;
Minsupp , minconf : threshold values ;
Lsupp, Lconf: local support, local confidence
M : number of data sources
N : number of rules

Output :

R : synthesized association rules

1. For each rule Ri in S do
   Num(Ri) the number of data sources that contain rule Ri in S ;
   $wRi \quad \frac{Num(Ri)}{\sum_{j=1}^{n} Num(Rj)}$
2. For each data source do
   $wDi \leftarrow \sum_{m} wRi$
3.. For each rule Ri in S do
   GSupp(Ri) $\leftarrow \sum_{I=1}^{m} wDi * Lsuppi(R)$
   Gconf(Ri) $\leftarrow \sum_{I=1}^{m} wDi * Lconfi(R)$
4. Rank all rules in S by their supports.
5. Output the high-rank rules in S whose support and confidence are at least minsupp and minconf respectively.

4.3 Impact

This synthesizing rules by weighting algorithm has been implemented in Java language (Jdk1.5) with Ms-Access and it runs on Intel based Personal Computers. The method of ranking valid rules using synthesize by weighting in this algorithm has only less cost. It does not involve any complicated formula computation. Hence the algorithm has less computation and time consumption than the Xindong method (Reference-2)whose time complexity is greater than O(n2).The time complexity of our algorithm is O(n2), where n is the variable number on which the rules are defined. The space complexity is also reduced.

V. CONCLUSION

The main advantages of this method are simple calculations and low error amount. The time complexity of our algorithm as obtained is O(n2) whereas the time complexity suggested by Xindong is higher than O(n2), moreover the space complexity is also optimized as we have removed the normalization step where weight of rule and frequency has been multiplied in the Xindong method. The higher global support of patient oriented variables show that greater emphasis should be on parameters like grievance handling, privacy and security in hospitals for improving service quality.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. Of Very Large Databases, Santiago, Chile,1994, pp. 487-499.

[2] R. Bayardo, "Efficiently mining long patterns from databases," in Proc. Of ACM SIGMOD Int'l. Conf .Management of Data, 1998, pp. 85-93, .

[3] K. Gouda and M. J. Zaki, "GenMax: An Efficient Algorithm for Mining Maximal Frequent Itemsets," Data Mining and Knowledge Discovery, 11(3):223-242,2005.

[4] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," in Proc. Of ACM SIGMOD Int'l. Conf. Management of Data, Dallas,2000 pp. 1-12.

[5] C.Lucchese, S. Orlando, R. Perego and C.Silvestri,"Mining Frequent Closed Itemsets from Distributed Repositories," Knowledge and Data Management in GRIDs, Part IV, pp. 221-234, 2007.

[6] C. Lucchese, S. Orlando, and R. Perego, "Fast and Memory Efficient Mining of Frequent Closed Itemsets,"IEEE Trans. Knowledge and Data Engineering, 18(1):21-36, January 2006.

[7] J. Pei, J. Han and R. Mao, "CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets," in Proc. of SIGMOD Int'l Workshop Data Mining and Knowledge Discovery, 2000, pp. 21-30.

[8] J.Wang, J. Han, Y. Lu, and P. Tzvetkov, "TFP: An Efficient Algorithm for Mining Top-K Frequent Closed Itemsets," IEEE Trans. Knowledge and Data Engineering,17(5):652-653, May 2005.

[9] J.Wang, J. Han, and J. Pei, "CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets," in Proc. Of ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Washington, 2003, pp. 236-245.

[10] M. Zaki and C. Hsiao, "Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure," IEEE Trans. Knowledge and data engineering,17(4):462-478, April 2005.