DISTRIBUTED DATA MINING FOR SYNTHESIZING HIGH FREQUENCY ASSOCIATION RULES : A CASE STUDY FOR DETERMINING SERVICE QUALITY IN HOSPITALS

Anirban Chakrabarty, Sonal G. Rawat

ABSTRACT

Many large organizations have multiple data sources, while putting all data together from different sources might amass a huge database for centralized processing. Data mining involves the exploration and analysis of large amounts of data in order to discover meaning patterns. Data mining association rules at different data sources and forwarding the rules to the centralized company headquarter provides a feasible way to deal with multiple data source problems. However, the forwarded rules from different data sources may be too many for the centralized company headquarter to use. Therefore, there is a need to find high frequency rules that have major role in decision making process.

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. 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, SERVQUAL scale.

1. 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 $A \rightarrow B$, where A, B $\subset$ I, and $A \cap B$ =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 statistical measure called support, denoted as supp. For an itemset $A \subset I$, supp(A) = s,if the fraction of transactions in D containing A equals to s. A rule $A \rightarrow B$ has a measure of strength called confidence (denoted as conf) which is defined as the ratio supp($A \cup B$) / supp(A).

The problem of mining association rules is to generate all rules $A \rightarrow B$ that have both support and confidence greater than or equal to some user specified threshold, called minimum support (minsupp) and minimum confidence (minconf), respectively.

For regular associations,

supp(AuB)≥ minsupp,

conf($A \rightarrow B$) = supp(AuB) / supp(A) ≥ minconf.

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 postmining 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.

Based on the above analysis, the problem can be formulated as follows: Given n data sources from a large organization, we are interested in mining each of these data sources for local rules for each data source synthesizing these local rules to find valid rules for the overall organization that would have been discovered from the union of all these data sources

There are various existing data mining algorithms that can be used to discover local rules for each data source. This work only focuses on synthesizing valid rules for the organization.

The proposed method for synthesizing valid rules has been applied for patient perceived service quality in hospitals. The remaining part of this paper is organized as follows: section 2 describes the problem of synthesizing valid rules with a case study of patient perceived service qualities in a hospital. Then section 3 presents both informal and formal description of algorithm respectively. Finally, in section 4, an analysis on performance is discussed.

2. PROBLEM DESCRIPTION

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 2 gives the local support and local confidence values for the rules RI, 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.

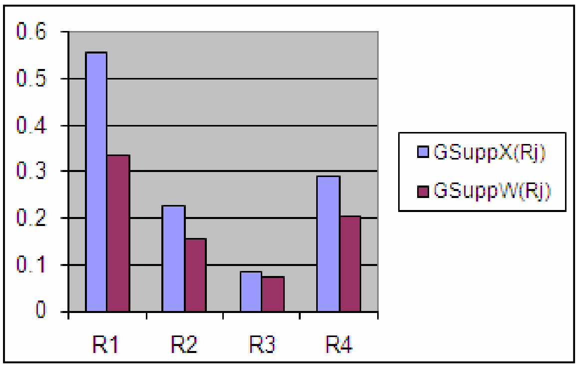

Fig. 1. Performance Analysis Bar Diagram

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 $\rightarrow$ Y, its global support and global confidence are defined as follows:

$$Gsupp(Ri): \leftarrow \sum_{i=l}^{m} wD_i * Lsupp(R_i)$$

$$Gconf(Ri): \leftarrow \sum_{i=l}^{m} wD_i * Lconfi(R_i)$$

where Gsupp(R_i) is the support of R_i after synthesizing, Gconf(R_i) 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.

Thus this paper devices a uniform framework for synthesizing valid rules as follows:

Phase 1 - Computing weight of rules.

Phase 2 - Computing weight of data sources.

Phase 3 - Computing global support and global confidence.

3. 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.

3.1 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. Table 2 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 :- S_1 = {R_1, R_2, R_4}

R_1 with Lsupp = 0.50, Lconf = 0.27

R_2 with Lsupp = 0.31, Lconf = 0.30

R4 with Lsupp = 0.47, Lconf = 0.82

Datasource D2:-

S_2 = {R_1, R_4, R_3}

R_1 with Lsupp = 0.40, Lconf= 0.69 ;

R_3 with Lsupp = 0.30, Lconf = 0.60 ;

R_4 with Lsupp = 0.27, Lconf = 0.59 ;

Data source D_3:- S_3 = {R_2, R_1}

R_1 with Lsupp = 0.43, Lconf = 0.73 ;

R_2 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:

wR_1 = 3/ (3+2+1+2) = 3/8 = 0.375 ;

wR_2= 2/(3+2+1+2) = 2/8 = 0..25 ;

wR_3= 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:

$$wR_i = \frac{Num(R_i)}{\sum_{j=1}^{n} Num(Rj)}$$

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.

wD_1 = (.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 R_1 : Patient oriented

$Gsupp(R_1) = wD_1 * Lsupp_1(R_1) + wD_2 * Lsupp_2(R_l) + wD_3 * Lsupp_3(R_l)$

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

=0.3354

$Gconf(R_1) = wD_1 * Lconf_1(R_1) + wD_2 * Lconf_2(Rl) + wD_3 * Lconf_3(Rl)$

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

For Rule R2: Technical Competence

$Gsupp(R_2) = wD_1*Lsupp_1(R2) + wD_3*Lsupp_3(R2)$

=0.2917* 0.31+ 0.2083 * 0.31

$=0.155$

$Gconf(R_2) = wD_1 * Lconf_1(R2) + wD3 * Lconf_3(R2)$

$=0.2917 * 0.30 + 0.2083 * 0.71$

$=0.2354$

For Rule R3 : Empathy

$Gsupp(R3) == wD_2 * Lsupp_2(R3)$

$= 0.25 * 0.30$

$= 0.075$

$Gconf(R_3) = wD2 * Lconf_2(R3)$

$= 0.25 * 0.60$

$= 0.15$

For rule R4 : Convenience

$Gsupp(R_4) = wD_1 * Lsupp_1(R_4) + wD2 * Lsupp_2(R4)$

$= 0.2917 * 0.47 + 0.25 * 0.27$

$= 0.2045$

$Gconf(R_4) = wD1 * Lconf_1(R_4) + wD_2 * Lconf_2(R_4) = 0.2917 * 0.82 + 0.25 * 0.59$

$= 0.3866$

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.

3.2 Formal Description

Algorithm :- Synthesizing Rules By Weighting method

Input:

$S = \{S_1, S_2, \ldots. Sm\}$: 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 ;

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

2. For each data source do

$$wDi \leftarrow \frac{\sum^{n} wRi}{m}$$

3.. For each rule Ri in S do

$$GSupp(Ri) \leftarrow \sum_{I=1}^{m} wDi * Lsupp_i(R)$$

$$Gconf(Ri) \leftarrow \sum_{I=1}^{m} wDi * Lconf_i(R)$$

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.

3.3 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 in our algorithm as we have removed the normalization step in calculating weight of data sources.

4. PERFORMANCE ISSUES

To evaluate the effectiveness, there are many possible measures one can choose to determine, how good our approach works. Here, we defined one type of error that is known as average error to measure the effectiveness of our approach compared with Xindong approach.

The formula for computing average error is given by

$$AE(R_i) = \frac{Gsupp^{x}(R_i) - Gsupp^{w}(R_i)}{Freq(R_i)} \quad i = 1,2 \ldots N$$

N : Total number of rules in the given set

Gsuppx(Ri): Golbal support of Ri in Xindong weighting model

Gsuppw(Ri): Golbal support of Ri in proposed weighting model.

Freq(Ri) : Frequency of Ri .

Table5 summarizes the result and gives the performance comparison of our algorithm with Xindong algorithm.The bar diagram showing Gsupport of the rules determined by Xindong method and our proposed method is given in Fig.1

The X-axis in graph represents the different rules in the data sources and the Y-axis represents the global support of these rules after synthesizing the rules.

5. CONCLUSION

Synthesizing rules by weighting model is presented in this paper in order to rank the rules which are getting from different branches of a hospital based on the patient perceived service qualities. This algorithm is a refinement work of Xindong weighting model. The performance analysis of this algorithm is also done with the result of less computational complexity.

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.

Table -1: Items of service quality scale

Variable Description :

V1 : **Hospital infection control** (HIC) Ability to reduce or eliminate infection risks to patients, visitors, and service providers in the hospital

V2: **Privacy** (PRI): The extent to which a hospital is able to maintain private record of patients

V3. **Grievance handling time** (GHT): Time taken by hospital administration to solve any grievances of patient

V4. **Continuity of care record** (CCR): Capacity of a hospital to maintain proper and detailed record of the patients' case history, number of visits made, etc

V5. **Waste disposal policy** (WDP): It denotes the policy of a hospital related to handling, storage, transportation, and disposal of hazardous materials.

V6. **Waiting time** (WT): This variable indicates the total time spent by a patient for fixing appointment as well as consulting with the doctor

V7. **Access** (ACS): Ability of a hospital to admit patients for whom it can provide services with its available resources

V8.**Administrative staff's attitude** (ASA): The administrative staff's behavior towards patients, visitors, and practitioners

V9. **Facilities availability** (FA): Availability of specialized departments and facilities in the hospital like burn care, skin care, outdoor etc.

V10. **Practitioner's attitude** (PRA): The attitude of practitioner's towards patient and visitors

V11. **Cost of medicine** (COM): Cost related to the medicines prescribed by the doctors in the hospital

V12.**Security in Hospital**(SH): Your hospital provides adequate security in terms of fire control, theft, controlling agitation etc.

Table 2: Lsupport and Lconfidence values of the Rules

D	Rules	Lsupport	Lconfidence
D	R 1	0.50	0.27
	R2	0.31	0.30
	R4	0.47	0.82

D	R1	0.40	0.69
	R4	0.27	0.59
	R3	0.30	0.60
D	R2	0.31	0.71
	R1	0.43	0.73

Table 3. Gsupport and Gconfidence Using Xindong Method

Rule	Frequency	wRi	$Gsupp^{W}(R_{J})$	$Gconf^{w}(R_{J})$
R1	3	0.143	0.3354	0.4033
R2	2	0.429	0.155	0.2354
R3	1	0.286	0.075	0.15
R4	2	0.143	0.2045	0.3866

Table 4. Gsupport and Gconfidence Using Proposed Method

Rulle	Frequency	wR_i	$Gsupp^{X}(R_{J})$	$Gconf^{x}(R_{J}$
R1	3	0.375	0.5551	0.6621
R2	2	0.25	0.2271	0.4698
R3	1	0.125	0.0859	0.2068
R4	2	0.25	0.2916	0.5096

Table-5 : Results of Performance Analysis

Rule	$AE(R_{J})$	$GsuppX(R_{J})$	$Gsupp^{W}(R_{J})$
R1	0.07	0.5551	0.3354
R2	0.04	0.2271	0.155
R3	0.01	0.0859	0.075
R4	0.04	0.2916	0.2045

REFERENCES:

1. Xindong Wu, Knowledge Discovery in Multiple Databases, Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, 2004.
2. Xindong Wu and Shichao Zhang, Synthesizing High-Frequency Rules from Different Data Sources, IEEE Transactions on Knowledge and Data Engineering, Vol.15, No.2, 353-367, 2003.

3. 3, Koushiki Choudhury, Avinandan Mukhmiee and Ashish Banerjee, Relationship Marketing Strategies and Patient Perceived Service Quality, CRM in Financial Services, 2001.
4. R.Agrawal and R,Srikant, Fast Algorithms for Mining Association Rules, Very Large Database, IEEE Transactions on Knowledge and Data Engineering,Vol.9, No.3, 143-167, 2001.
5. 5.Abraham Silberchatz, Henry F.Korth and S.Sudarshan, Database System Concepts, Third Edition, McGraw-Hill International Edition, 1997.
6. J.Han, J.Pei and Y.Yin, Mining Frequent Patterns without Candidate Generation, ACM-SIGMOD, 1-12, 2000.
7. Matignon Randall, Data Mining Using SAS Enterprise Miner, Wiley Publications, 2008
8. Zdravko Markov, Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage, Wiley, 2008

Annexure-I

QUESTIONNAIRE

(For academic purpose only)

Name ____

Occupation ______________________________________

☐ Service ☐ Business ☐ Other, please specify

For each of the following statements, could you please tell me if you agree or disagree (YES/NO).

1. The procedure for admission in the hospital is not complicated and time consuming a. Yes b. No
2. The hospital has well qualified and efficient doctors in its panel. a. Yes b. No
3. The attitude of staff and administration towards patient is satisfying a. Yes b. No
4. Problems or grievances of patients are handled on time a. Yes b. No
5. Hospital does not charge any unfair claims from patients a. Yes b. No
6. Adequate support staff is present in the hospital a. Yes b. No
7. The hospital specializes in a particular treatment (a) Yes (b) No
8. Adequate surgical equipments and medicines are present in the hospital (a) Yes (b) No
9. The hospital has a proper waste disposal policy i.e. policy of a hospital related to handling, storage, transportation, and disposal of hazardous materials (a) Yes (b) No
11. Cost related to the medicines prescribed by the doctors in the hospital is not very high (a) Yes (b) No
12. The hospital has a proper chemist and druggist store (a)Yes (b) No
13. 12. Your hospital provides adequate security in terms of fire control, theft, controlling agitation (a)Yes (b) No

13. The hospital maintains detailed record of the patients' case history, number of visits made (a)Yes (b) No Hospital has proper arrangement of basic facilities like canteen, toilet, water and air conditions (a)Yes (b) No
14. Privacy of patients is given due importance by the hospital (a)Yes (b) No
15. The hospital's physical environment and ambience is satisfactory
(a)Yes (b) No
16. The in patient food service quality is satisfying. (a)Yes (b) No
17. Doctors and nurses provide personalized care to patients. (a)Yes (b) No
18. The charges of rooms and service in the hospital are economical
(a)Yes (b) No
19. There is separate department for infectious and fatal diseases. (a)Yes (b) No
20. The hospital has proper facilities for OPD and emergency circumstances.
(a)Yes (b) No
21. Doctors associated with the hospital are available on call (a)Yes (b) No
22. The hospital provides for good arrangement of staying for acquaintances of admitted patients (a)Yes (b) No
23. The hospital has a proper enquiry department to handle queries of patients.
(a)Yes (b) No
24. The surrounding of the hospital is healthy and hygienic and not any industrial area
(a) Yes (b) No
25. There are separate pathology and radiology departments in the hospital
(a) Yes (b) No
26. Hospital provides for corporate discounts (a) Yes (b) No
27. Hospital organizes for camps and subsidized health packages (a) Yes (b) No
28. A patient can conveniently fix an appointment for consulting with the doctor.
(a) Yes (b) No