A Model for Mining Course Information using Vague Association Rule

Anjana Pandey
UIT RGPV Bhopal

K.R.Pardasani
MANIT Bhopal

ABSTRACT

There are different university offering different types of courses over several years, and the biggest issue with that is how to get information to make course more effective. Association rule mining can be used to evaluate the course effectiveness and helps to look for in regards to changes in performance of the course. For Example there is a course offering different topics. We can say that the topics having full attendance are totally effective and carry no hesitation information. While there are some topics which are almost fully attendant carry some hesitation information. This hesitation information is valuable and can be used to make the course more effective and interesting. We use vague association rule to render that hesitation information and develop an algorithm to mine the hesitation information. Our experiments on real datasets verify that our algorithm to mine the Vague Association Rule is efficient. Compared with the traditional Association Rule mined from transactional databases, the Vague Association Rule mined from the AH-pair databases are more specific and are able to capture richer information.

Keywords

Hesitation Information,Vague Association Rule,AH pair.

1. INTRODUCTION

Consider the classical market basket case, in which Association Rule(AR) mining is conducted on transactions that consist of items bought by customers. There are many items that are not bought but customers may have considered to buy them. We call such information on a customer’s consideration to buy an item the hesitation information [1] of the item, since the customer hesitates to buy it. The hesitation information of an item is useful knowledge for boosting the sales of the item. However, such information has not been considered in traditional AR mining due to the difficulty to collect the relevant data in the past. Nevertheless, with the advances in technology of data dissemination, it is now much easier for such data collection.

A typical example is an online shopping scenario, such as “Amazon.com”, for which it is possible to collect huge amount of data from the Web log that can be modeled to mine hesitation information. From Web logs, we can infer a customer’s browsing pattern in a trail, say how many times and how much time s/he spends on a Web page, at which steps s/he quits the browsing, what and how many items are put in the basket when a trail ends, and so on. Therefore, we can further identify and Categorize different browsing patterns into different hesitation information with respect to different applications. The hesitation information can then be used to design and implement selling strategies that can potentially turn those “interesting” items into “under consideration” items and “under consideration” items into “sold” From the literature [1], it is evident that very little attention has been paid for mining hesitation information .In this paper an attempt has been made to develop a vague set model for mining hesitation information .It is illustrated with the help of problem of choosing a course in an educational institute.

There are many different type of status of a piece of hesitation information (called hesitation status (HS)) [2]. Let us consider an example of class scenario that involves following type of status: (s1) attended class between 0 - 20%; (s2) Attended class between 0-40% (s3) Attended class between 0-60%.All of the above-mentioned types of HS are the hesitation information of those classes. Some of the types of HS are comparable based on some criterion, which means we can define an order on these types of HSs. For example, given a criterion as the possibility that the student attended the classes, we have $S_1 \le S_2 \le S_3$ .Here we are employ the vague set theory [3,4,5] to model the hesitation status of the course attended by the students. The main benefit of this approach is that the theory addresses the drawback of a single membership value in fuzzy set theory [6] by using interval-based membership that captures three types of evidence with respect to an object in a universe of discourse: support, against and hesitation. Thus, we naturally model the hesitation information of a course in the mining context as the evidence of hesitation.

The information of the “attended the class” and the “not attended the class” (without any hesitation information) in the traditional setting of association rule mining correspond to the evidence of support and against with respect to the class.

To study the relationship between the support evidence and the hesitation evidence with respect to topics, the concepts of attractiveness and hesitation are used, which are derived from the vague membership in vague sets. A topic with high attractiveness means that the topic is well attended and has a high possibility to be attended again next time. A topic with high hesitation means that the student is always hesitating to attend the topic due to some reason but has a high possibility to attend it next time if the reason is identified and resolved. For example, given the vague membership value, [0.5, 0.7], of a topic, the attractiveness is 0.6 (the median of 0.5 and 0.7) and the hesitation is 0.2 (the difference between 0.7 and 0.5), which implies that the student may attend the topic next time with a possibility of 60% and hesitate to attend the topic with a possibility of 20%.Using the attractiveness and hesitation of topics, we model a database with hesitation information as an AH-pair[4] database that consists of AH-pair transactions, where A stands for attractiveness and H stands for hesitation. Based on the AH-pair database, we then employed the notion of Vague Association Rules, which capture four types of relationships between two sets of items: the implication of the attractiveness/ hesitation of one set of items on the attractiveness/hesitation of the other set of items. For example, if we find an AH-rule like "People always buy quilts and pillows (A) but quit the process of buying beds at the step of choosing delivery method (H)". Thus, there might be something wrong with the delivery method for beds (for example, no home delivery service provided) which causes people hesitate to buy beds. To evaluate the quality of the different types of Vague Association Rule, four types of support and confidence are defined. We also investigate the properties of the support and confidence of Vague Association Rule, which can be used to speed up the mining process.

This paper is organized as follows. Section 2 gives some preliminaries on vague set and association rules. Section 3 discusses the algorithm that mines vague association rules. Section 4 illustrates the example. Section 5 reports the experimental results. Section 6 concludes the paper.

2. PRELIMINARIES

The following definitions have been used to develop the model and algorithm for mining vague association rules.

2.1 Vague Sets

Let I be a classical set of objects, called the universe of discourse, where an element of I is denoted by x .

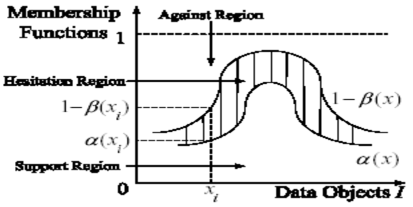

Fig1The true (α) and false (β) Membership functions of a vague Set

A vague set V in a universe of discourse I is characterized by a true membership function, α_V , and a false membership function, β_V as follows: $\alpha_V : I \rightarrow [0,1]$, $\beta_V : I \rightarrow [0,1]$,where $\alpha_V(x)+\beta_V(x) \leq 1$, $\alpha_V(x)$ is a lower bound on the grade of membership of x derived from the evidence for x , and $\beta_V(x)$ is a lower bound on the grade of membership of the negation of x derived from the evidence against x .Suppose $I = \{x_1, x_2, ... x_n\}$ .A vague set V of the universe of discourse I is represented $V = \sum_{i=1}^{n} [\alpha(x_i), 1-\beta(x_i)]/x_i$,where $0 \leq \alpha(x_i) \leq (1-\beta(x_i))) \leq 1.$

The grade of membership of x is bounded to $[\alpha(x), 1-\beta(x)]$,which is subinterval of $[0,1]$ fig 1.Here $[\alpha(x), 1-\beta(x)]/x$ is a vague element and the interval $[\alpha(x), 1-\beta(x)]$ is the vague value of the object x . For example, $[\alpha(x), 1-\beta(x)]$ = [0.5, 0.7] is interpreted as "the degree that the object x belongs to the vague set V is 0.5 (i.e. α(x)) and the degree that x does not belong to V is 0.3 (i.e. β(x))." For instance, in a voting process, the vague value [0.5, 0.7] can be interpreted as "50% of the votes support the motion, 30% are against, while 20% are neutral (abstentions)."[5]

2.2Median Membership and Imprecision Membership

To compare vague values, we use two derived memberships: median membership and imprecision membership [5]. We have unique median membership $M_m(x)$ and imprecision membership $M_i(x)$, for a given vague value $[\alpha(x), 1-\beta(x)]$ .Median membership is defined as,

$$M_m = \frac{1}{2}(\alpha + (1-\beta))$$

which represents the overall evidence contained in a vague value. It can be checked that $0 \leq M_m \leq 1$.The vague value [1, 1] has the highest M_m, which means the corresponding object definitely belongs to the vague set (i.e., a crisp value). While the vague value [0, 0] has the lowest M_m , this means that the corresponding object definitely does not belong to the vague set.Imprecision membership is defined as $M_i = ((1-\beta)-\alpha)$, which represents the overall imprecision of a vague value. It can be checked that $0 \leq M_i \leq 1$. The vague value $[a,a](a \in [0,1])$ has the lowest M_i which means that the membership of the corresponding object is exact (i.e., a fuzzy value). While the vague value [0, 1] has the highest M_i which means that we do not have any information about the membership of the corresponding object. The median membership and the imprecision membership are employed to measure the attractiveness and the hesitation of a topic with respect to a student.

A hesitation status (HS) is a specific state between two certain situations of "attending" (100% classes) and "not attending" (0 % classes) the class of the particular topic of a course. The hesitation information is formally defined as follows.

2.3 Hesitation and Overall Hesitation

Given an item $x \in I$ and a set of HS $S=\{S_1, S_2 .... S_w\}$ with a partial order $\leq$.The hesitation of x with respect to an HS $s_i \in S$ is a function $h_i(x) : I \rightarrow [0,1]$,such that $\alpha(x)+\beta(x)+\sum_{i=1}^{w} h_i(x) = 1$, where $h_i(x)$ represent the evidence for the HS s_i of x.The overall hesitation of x with respect to S is given by $H(x) = \sum_{i=1}^{w} h_i(x)$ [3].

2.4 Intent and Overall Intent

Given a set of HS, $(S, \leq)$, the intent of an item x with respect to HS $s_i \in S$,denoted as $\text{int}(x, s_i)$,is a vague value $[\alpha_i(x), 1-\beta_i(x)]$ which is a subinterval of $[\alpha(x), 1-\beta(x)]$

and satisfy the following conditions

1. (1-$\beta_i(x)$)= $\alpha_i(x)+h_i(x)$
2. if s_i is in chain of the CG, $s_1 \leq s_2 \leq .... \leq s_i$,then for $1 \leq i \leq n$ we define The overall intent of x ,denoted as $INT(x)$, is the interval $[\alpha(x), 1-\beta(x)]$ [3].

2.5 Attractiveness and overall Attractiveness

The attractiveness of X with respect to an HS S_i ,denoted as $att(x, S_i)$ is defined as the median membership of X with respect to S_i ,that is $\frac{1}{2}(\alpha_i(x)+(1-\beta_i(x)))$

The overall attractiveness[7] of X is a function $ATT(x): I \rightarrow [0,1]$,such that $ATT(x)=\frac{1}{2}(\alpha(x)+(1-\beta(x)))$

2.6 AH- pair Transaction and database

An AH-pair transaction T is a tuple $< v_1, v_2, ..., v_m >$ on an itemsets $I_T = \{x_1, x_2, ... x_m\}$ where $I_T \subseteq I$ and $v_j =< M_A(x_j), M_H(x_j) >$ is an AH pair of the item x_j with respect to a given HS or the overall hesitation, for $1 \le j \le m$ .An AH-pair database is a sequence of AH-pair transactions [7].

2.7 Vague Association Rule

A Vague Association Rule (VAR) $r = (X \Rightarrow Y)$,is an association rule obtained from an AH-pair database.There are four types of support and confidence to evaluate the VARs as follows

2.8 Support

Given an AH- pair database, D, we define four types of support for an itemset Z or a VAR $X \Rightarrow Y$,where $X \cup Y = Z$ as follows [8].

1.The A- support of Z ,denoted as $A\sup p(Z)$,is defined as $\frac{\sum\limits_{T \in D} \prod\limits_{z \in Z} M_A(z)}{|D|}$

2.The H-support of Z ,denoted as $H\sup p(Z)$,is defined as $\frac{\sum\limits_{T \in D} \prod\limits_{z \in Z} M_H(z)}{|D|}$

3.The AH -support of Z ,denoted as $AH\sup p(Z)$,is defined as $\frac{\sum\limits_{T \in D} \prod\limits_{x \in X, y \in Y} M_A(x) M_H(y)}{|D|}$.

4. The HA-support of Z ,denoted as $HA\sup p(Z)$,is defined as $\frac{\sum\limits_{T \in D} \prod\limits_{x \in X, y \in Y} M_H(x) M_A(y)}{|D|}$. Z is an A (or H or AH or HA) FI if the A - or H or AH or HA) support of Z is no less than the (respective A or H or AH or HA) minimum support threshold where FI means frequent itemsets.

2.9 Confidence

Given an AH- pair database, D, we define four types of support for an itemset Z or a VAR $X \Rightarrow Y$,where $X \cup Y = Z$ as follows[8].

1. If both X and Y are A FIs, then the confidence of r , called the A-confidence of r and denoted as $Aconf(r)$, is defined as . $\frac{A\sup p(Z)}{A\sup p(X)}$

2. If both X and Y are H FIs, then the confidence of r , called the H-confidence of r and denoted as $Hconf(r)$, is defined as . $\frac{H\sup p(Z)}{H\sup p(X)}$

3. If X is an A FI and Y is an H FIs, then the confidence of r , called the AH confidence of r and denoted as $AHconf(r)$, is defined as . $\frac{AH\sup p(Z)}{H\sup p(X)}$

$$ATT(x)=\frac{1}{2}(\alpha(x)+(1-\beta(x)))$$

$$\frac{HA\sup p(Z)}{H\sup p(X)}$$

4. If X is an $_H$ FI and Y is an A FIs, then the confidence of r , called the HA confidence of r and denoted as $HAconf(r)$, is defined as

3. ALGOITHM FOR MINING FOR VAGUE ASSOCIATION RULE

We present an algorithm to mine Vague Association Rules. We first mine the set of all A, H, AH and HA FIs from the input AH pair database with respect to certain HS or the overall hesitation. Then, we generate the Vague Association Rules from the set of FIs.To generate the A, H, AH and HA pair from the database first module is developed to calculate the Intent of an item .The intent of an item X, denoted as intent(x), is a vague value [α(x), 1 − β(x)]. The vague value of intent is calculated using the Algorithm **CalIntent().**

The calIntent() Algorithm which is first module is a nested iterative method to calculate the intent. This algorithm takes a Data-set (D) as input as given in the Table 1. This Data-set consists of rows and column as student ID (S_ID) and topic ID (T_ID) of the course. Therefore, data set D is considered as a two dimensional array.Step 1 initializes the intent array (having size as no. of topics) while Step 2 and Step 4 are used to navigate in the Data-set array. In Step 3 favor (α) and against (β) are initialized to store overall favor and against which is finally stored in the intent array in the Step 8. This algorithm3.1 returns an intent array as shown in Table 2.

3.1 Algorithm CalIntent(D)

1. Initialize intent array to store intent;
2. For each i=0,1,2…..where i<no. of tpID, do
3. Initialize favor(α) & against(β) variable with value zero;
4. For each j=0,1,2…..where j<no. of stID, do
5. Increment favor(α) by one when D[i][j] is equal to one;
6. Increment against(β) by one when D[i][j] is equal to zero;
7. End of for ;
8. Generate intent using favor and against as [α,1-β] ;
9. End of for;
10. return all intent;

The C**alAHPair** Algorithm 3.2 which is the second module is a simple iterative method to calculate the AH pair. This algorithm takes a Intent as input as given by algorithm 3.1.Step 1 initialize the AH pair array having size as no. of tpID. Step 2 is used to traverse the intent array while Step 3, 4, 5 are used to calculate attractiveness and hesitation to finally calculate the AH pair. This algorithm returns AH -pair array as shown in Table 3.

3.2 Algorithm CalAHPair

```
1. Initialize AHPair array to store AH pair;
2.For each i=0,1,2……..where i<no. of tpID
3.Attractiveness as a median membership i.e. ½(α+(1-β));
4.Hesitation as a difference of α and 1-β using intent;
5.End of for;
6.return all AHPair;
```

Now we have AH pair database D to generate vague association rules. Let A_i and H_i be the set of A frequent itemset (FIs) and H frequent itemsets containing i items, respectively. Let $A_i H_j$ be the set of AH frequent itemsets containing i items with A values and j items withH values. Here $A_i H_j$ is equivalent to $H_j A_i$. Let C_W be the set of candidate FIs, from which the set of FIs W is to be generated, where W is A_i, H_i, or $A_i H_j$.

3.3Algorithm MineVFI (D, σ)

```
1. Initialize FIs array to store FI;
2. Mine $A_1$ and $H_1$ from $D$;
3. Generate $C_{A_2}$ from $A_1$, $C_{A_1H_1}$ from $A_1$ and $H_1$,and $C_{H_2}$ from $H_1$;
4.Calculate the support of $C_{A_2}, C_{A_1H_1}$ and to give $A_2, A_1H_1$ and $H_2$,respectively ;
5. for each $k = 3,4,.......,$ where $k = i + j, do$
6.Generate $C_{A_k}$ from $A_{i-1}$ and $C_{H_k}$ from $H_{i-1}$,for $i = k$ ;
7.Generate $C_{A_iH_j}$ from $A_{i-1} H_j$,for $2 \le i < k$ and from $A_1 H_{j-1}$,for $i = 1$;
8.Calculate the support of $C_{A_k}, C_{A_iH_j}$ and $C_{H_k}$ to give $A_k, A_i H_j$ and $H_k$,respectively ;
9 If all $A$ ,$H$, $AH$, are greater than $\sigma$ add into the array $FIs$ ;
10 return all $FIs$;
```

The algorithm to compute the frequent itemset is shown in Algorithm 3.3. We first mine the set of frequent items A_1 and H_1 from the input AH-pair database D. Next, we generate the candidate FIs that consists of two items (Line 2) and compute the FIs from the candidate frequent itemset (Line 3). Then, we use the frequent itemset containing $(k-1)$ items to generate the candidate frequent itemset containing k items, for $k \ge 3$.The support of the candidate frequent itemset is computed and only those with support at least σ are retained as frequent itemset . Finally, the algorithm terminates when no candidate frequent itemset (FIs) are generated and returns all FIs.

After mining the set of all frequent itemset s, Vague association rule are generated from the frequent itemset. There are four types of Vague association rule. First, for each A or H FI Z, we can generate the vague association rule $X \Rightarrow Y$, $\forall X, Y$ where $X \cup Y = Z$ using the classical association rule generation algorithm [9]. Then, for each AH (or HA) frequent itemset $Z = X \cup Y$, where X is an A frequent itemset and Y is an H frequent itemset , we generate two Vague Association Rule $X \Rightarrow Y$ and $Y \Rightarrow X$. The confidence of the vague association rule can be computed by Definition 2.9.

4. ILLUSTRATION OF ALGORITHM

Table 1 shows the data of student, where 1 and 0 represents that student attends the class and student does not attended the class (without any hesitation information) respectively as in traditional association rule mining setting. The set of hesitation status is given by $S = \{S_1, S_2, S_3, S_4, S_5\}$ .The table 1 is constructed using data regarding attendance of course data structure taught in UIT RGPV Bhopal

Table 1 Sample Database of Attendance

S_ID	T_ID=1	T_ID=2	T_ID=3	T_ID=4
1	1	S_4	S_4	S_1
2 C_{H_2}	1	0	S_1	0
3	1	1	S_3	S_3
4	0	S_5	S_2	S_3
5	S_1	1	S_5	S_2
6	1	0	S_3	S_5
7	1	S_5	S_4	0
8	S_1	0	0	S_2
9	S_3	0	1	0
10	1	S_5	0	S_5

First we calculate the intent of topic with respect to S_1.

Table 2 Intent of Topic with respect to S_1

T_ID=1	T_ID=2	T_ID=3	T_ID=4
[0.6,0.9]	[0.2,0.6]	[0.1,0.8]	[0.0,0.7]

The intent database of all topic (1,2,3,4) for different HSs (S_2, S_3, S_4, S_5) can be similarly determined. In next step we calculate AH pair by algorithm 3.2 where input is intent.

Table 3 AH –pair database

A	*H*	*AH*	*HA*
[0.3375]	[0.2099]	[0.1349]	[0.525]

Now we calculate the support of $C \Rightarrow A$

T_ID=1	T_ID=2	T_ID=3	T_ID=4
[0.7500,0.2999]	[0.4000,0.4000]	[0.4500,0.7000]	[0.3500,0.7000]

Table 4 support of $C \Rightarrow A$

Table 5 The support of $B \Rightarrow A$

A	H	AH	HA
[0.3]	[0.1199]	[0.11]	[0.3]

5. PERFORMANCE VULATION AND EXPERIMENT

To present the scale-up properties of our algorithms, we performed experiments on a 2.4GHz, 512 Mb PC running Windows XP Professional. The algorithm is implemented in Java. For the experiment, the data of attendance of the Data-structure Course, taught in Information Technology of UIT-RGPV Bhopal is used to prepare the Data-set (D) according to the format illustrated in the table 1 as a array in which column contains the Hesitation status of different topics (introduction to data structure, array, stacks, queue …..) and rows contains student IDs. The trends discovered are aggregated to finally make conclusion. We identify many trails on the data-set and aggregated them to finally come to a conclusion.When σ=0.002 we obtain many vague association rules some of them are as given.

1. $Array \Rightarrow Stack$ With HA support =0.78.

Rule 1 shows that topic Array is prerequisite for topic Stack .Topic of Stack heavily depended on Array topic. If the students have no knowledge of Arrays then students show more hesitation to attend the lectures on Stacks.

2. $Stack \Rightarrow Queue$ With HA support = 0.10

Rule 2 illustrates that topic of Stack is not prerequisite for the topic Queue and there is little dependency among them. So students can directly attend the lecture of Queue with less hesitation.

3. Array, Tree => Graph with HA support = 0.42

Rule 3 shows that the Introduction of Graph depended on the topic Array and queue both with the HA support 0.42 which finally concluded that when the topic Graph are taught without introduction of Array and Tree then student feels hesitation to attend that course .

Fig.2 and Fig. 3 report the running time and the number of FIs. From Fig. 2, the running time decreases with the Increase in the value of σ due to the larger number of FIs generated. Fig. 3 shows that the number of FIs decrease with the support increases.

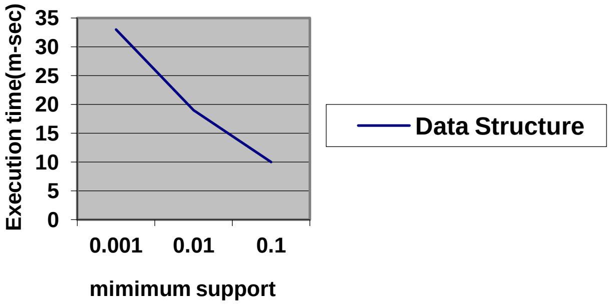

Fig.2 Running Time of Frequent itemset

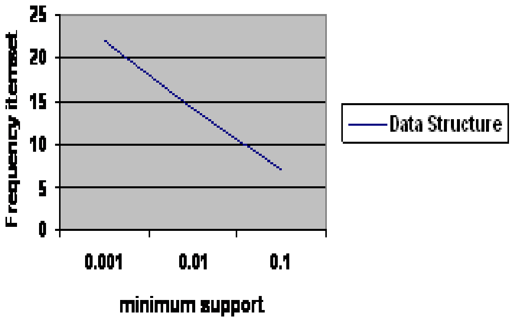

Fig.3 Number of Frequent Itemset

6 CONCLUSION

The models for hesitation information is developed by vague set theory in order to address a limitation in traditional association rule mining problem, which ignores the hesitation information of items in transactions. The efficient algorithm for mining vague association rule that discovers the hesitation information of items is proposed. The effectiveness of algorithm is also revealed by experiments on real datasets. This algorithm has wide applications for example different ranking scores together with click through data of a search result can be modeled as an object having different hesitation status. In this case vague association rule can be used to reflect different users' preferences. Such models can further be developed and extended to problems involving mining of hesitation information in different conditions. Also this algorithm is extended for mining temporal association rule.

7. REFERENCES

[1] An Lu and Wilfred Ng "Maintaining consistency of vague databases using data dependencies"Data and Knowledge Engineering,Volume 68,2009,Pages 622-641.

[2] Lu, A., Ng,W "Managing merged data by vague functional dependencies". In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. (eds.) ER 2004. LNCS, vol. 3288, pp. 259–272. Springer, Heidelberg.

[3] An Lu and Wilfred Ng "Mining Hesitation Information by Vague Association Rules"Lecture Notes in Computer Science ,Springer Volume 4801/,2008,pg 39-55.

[4] Gau, W.-L., Buehrer, D.J."Vague sets". IEEE Transactions on Systems, Man, and Cybernetics 23(2),1993, 610–614 .

[5] Lu, A., Ng, W."Vague sets or intuitionistic fuzzy sets for handling vague data": Which one is better? In: Delcambre, L.M.L., Kop, C., Mayr, H.C., Mylopoulos, J., Pastor, ´ O. (eds.) ER 2005. LNCS, vol. 3716, pp. 401–416. Springer, Heidelberg.

[6] Zadeh L. A.,"Fuzzy sets," *Inform. Contr.*, vol. 8, 1965,pp. 338–353.

[7] Lu.A.,Ng.,W:Handling Inconsistency of vague relations with functional dependencies.In :ER(2007).

[8] Lu,A.,Ke,Y.,Cheng ,J.,Ng,W.:Mining Vague association rules.In:DASFAA,pp.891-897(2007)

[9] Agrawal.R and Srikant.R. Fast algorithms for mining association rules. In Proc. of the 20th Int'l Conference on Very Large Databases,1994 Santiago, Chile.