

Enhancement of Marketing Strategies using Weighted Association Rule Mining

Dhwaneel Trivedi
Student of B.E (Computer)
TCET(Mumbai University)
Mumbai, India

Suraj Singh
Student of B.E (Computer)
TCET(Mumbai University)
Mumbai, India

Rashmi Thakur
Asst.Professor at TCET
TCET(Mumbai University)
Mumbai, India

ABSTRACT

This paper symbolizes our aim to promote Business organizations to take marketing decisions based on mining large databases of Transactions. Frequent item sets and Strong Association Rules are formed without need of supplying minimum support and minimum confidence. We have proposed and implemented an algorithm that scans Database only once and modifies Apriori algorithm and produces better results.

General Terms

Algorithms.

Keywords

Support, Confidence, weighted support, weighted confidence.

1. INTRODUCTION

Data mining [2] is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. Marketers can use this Algorithm specifically for Market Basket Analysis. It also enables various companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

The paper[5] motivated us because unlike a general apriori Algorithm[10] which considers each transaction equally, this algorithm proposed will provide weights on transactions by introducing concept of HITS (Hyperlink Induced Text Search) and will use apriori algorithm on w-support (weighted support) and w-confidence (weighted confidence). Hence it performs weighted association rule mining without pre-assigned weights.

The specialty of this algorithm is that it generates minimum support and minimum confidence from weighted supports and gives a choice to the Marketers to set a threshold value to get desired results. Hence they are not to supply values of minimum support and minimum confidence whenever they run this Algorithm.

This Algorithm does not treat all transactions equally and hence treats Transactions containing items as Hubs and individual items as Authorities. Hence HITS (Hypertext Induced Text Search) [3] is applied on the database and Transactions are graded. So, obviously the Transactions with higher weights contain more significant items and also more significant items are contained in highly weighted Transactions [1].

The basic idea behind w-support is that a frequent item set may not be important as it appears, because the weights of transactions are different. These weights are completely derived from the internal structure of the database based on the assumption that good transactions consist of good items.

This assumption is exploited by extending Kleinberg's HITS model and algorithm to bipartite graphs. Therefore, w-support is distinct from weighted support in weighted association rule mining (WARM) [6] [9], where item weights are assigned. Furthermore, a new measurement framework of association rules based on w-support [7] is proposed. This framework also generates minimum confidence from w-support and generates rules accordingly.

So, various interesting patterns can be retrieved from database based on inherent associations between transactions and items. A Marketer can change the threshold minimum support value from proposed values and observe variations in the results.

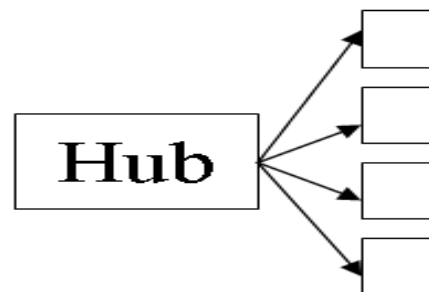


Fig 1: Hub containing Authorities

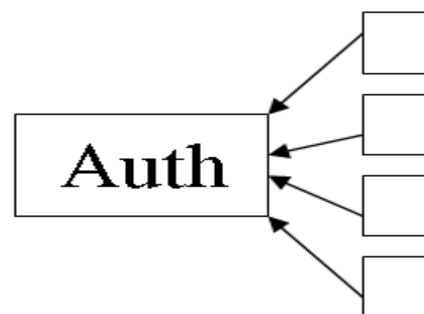


Fig 2: Authorities contained in Hubs

2. LITERATURE REVIEW

Association rule mining [4] is one of the important problems in data mining. The goal of the Association rule mining is to detect relationships or associations between specific values of categorical variables in large data sets. This is a common task in many data mining projects. Suppose I is a set of items, D is

a set of transactions, an association rule is an implication of the form $X \Rightarrow Y$, where X, Y are subsets of I , and X, Y do not intersect. Each rule has two measures, support and confidence. The association rule mining on Market "Basket Data" is Boolean Association Rule Mining (ARM) in which only Boolean attributes are considered. In order to do association rule mining on quantitative data, such as Remotely Sensing Image data, some mapping should be done from quantitative data to Boolean data. The main idea here is to partition the attribute values into Transaction Patterns. Basically, this technique enables analysts and researchers to uncover hidden patterns in large data sets.

Data mining or knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too Time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include:

- Market segmentation - Identify the common characteristics of customers who buy the same products from your company.
- Customer churn - Predict which customers are likely to leave your company and go to competitor.
- Fraud detection - Identify which transactions are most likely to be fraudulent.
- Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.
- Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
- Market basket analysis - Understand what products or services are commonly purchased together; e.g., Eggs and Milk.
- Trend analysis - Reveal the difference in typical customers between current and last months.

In our implemented Algorithm we direct ourselves towards identifying Frequent Item sets and Association Rules concentrating on Market Basket Analysis.

3. PROCEDURAL CONCEPT

Let T denote Transactions, D denote Database then the relationship between hubs and authorities is given by the equations:

$$\text{auth}(i) = \sum_{T: i \in T} \text{hub}(T)$$

$$\text{hub}(T) = \sum_{i: i \in T} \text{auth}(i)$$

The w-support of an item set X is defined as

$$\text{wsupp}(X) = \frac{\sum_{T: X \subseteq T} \text{hub}(T)}{\sum_{T \in D} \text{hub}(T)}$$

The w-support of an association rule $X \Rightarrow Y$ is defined as $\text{wsupp}(X \Rightarrow Y) = \text{wsupp}(X \cup Y)$.

The w-confidence is given by $\text{wconf}(X \Rightarrow Y) = \text{wsupp}(X \cup Y) / \text{wsupp}(X)$.

The w-confidence can be understood as the ratio of the hub weights received by X together with Y to the total hub weights received by X . Basically, w-support measures how significantly X and Y appear together; w-confidence measures how strong the rule is. If $\text{wconf}(X \Rightarrow Y)$ is large, it shows that many good hubs that vote X also vote Y , although the fraction of these hubs may be small.

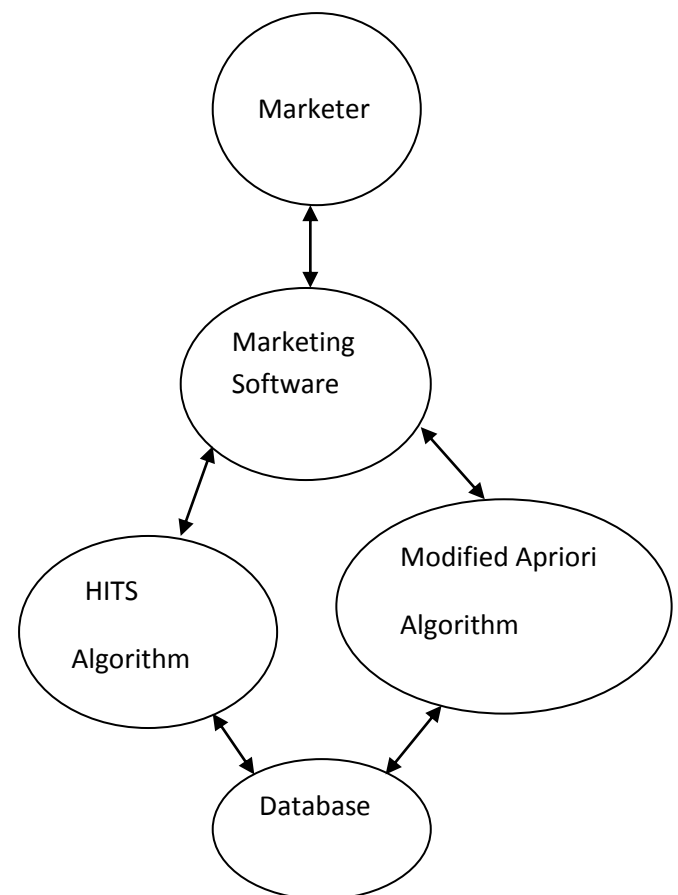


Fig 3: Implementation of Concept

4. ALGORITHM

The Algorithm described below has been successfully implemented using C# as a programming language in Microsoft Visual Studio 2010 as front end and Microsoft SQL Server 2008 as backend.

The Algorithm initially scans the database to create a Data Table(C#) from a dynamically generated table in database. Then this table is deleted from database. It consists of attributes having binary values as '1' or '0' denoting presence and absence of an item in a transaction respectively. The first column belongs to Transaction id and other columns belong to

items. Throughout the algorithm all operations are performed on Data table rather than database and hence the response time will be reduced drastically.

For implementing the above idea we initially generate a dynamic table by executing a query "create table trans1_dynamic (T_id varchar(10))". Then we insert Transaction id column from Transactions_table in database by "insert into trans1_dynamic(T_id) select distinct T_id from Transactions_table". The Transactions_table contains only two columns Transaction id and item id. Then we create temporary Data Table for storing item ids by "Select distinct item_id from Transactions_table" and then we fill temporary Data Table. Then for all items names we add columns in trans1_dynamic by "alter table trans1_dynamic add "+name+" int not null default(0)". Now we create fill final Data Table final_table from trans1_dynamic by "Select * from trans1_dynamic". Now we have to enter "1" in places where item is present in a corresponding Transaction. This can be done by creating a temporary Data Table trans_table and filling it by "Select * from Transactions_table". Let i=0, j=0.

While (i<number of rows in trans_table) {

While (j<number of rows in trans_table and Transaction ids of trans_table and final_table match){

Extract item_id from trans_table and make an entry of 1 in final_table at that item_id index,

Increment j.}

Increment i.}

Then we delete the trans_dynamic table from Database. Hence we have transformed a two column Table in database to Multi-Dimensional Data table final_table which would be treated as an array for implementing HITS and Modified Apriori algorithm.

Now we apply HITS algorithm on final_table to calculate hub weights and Authority weights.

4.1 HITS

For all items i initialize auth (i) =1

For (l=0; l< num_it ; l++) do begin

For each i set auth' (i) =0

For all transaction t belongs to Data Table do begin

Hub (t) = sum of all auth (i) where i belongs to t
 Auth' (i) +=hub (t) for each item i belongs to t

End

Auth (i) =auth' (i) for each item i belonging to t,
 Normalize auth

End

4.2 Calculation for weighted support:

for (j = 1; j < num_col ; j++) do begin
 initialize sum to 0;
 for (i = 0; i < num_rows; i++) do begin
 sum = addition of hub[i]
 end
 weighted_support[j] = sum;
 end

Initialize total to 0;
 for (i = 0; i < num_rows; i++) do begin
 total += hub_weight[i];
 end

for (i = 1; i < weighted_support.Length ; i++) do begin
 weighted_support[i] /= total;
 end

Then we apply Modified Apriori Algorithm where minimum support (minsupp) is generated from hub weight (hub) weighted support (wsupp) by taking average of weighted supports of items. If the Marketer wants to increase number of items in the frequent item sets he can decrease minimum support exponentially [8].

4.3 Modified Apriori Algorithm:

L1= { {i} | wsupp(i) > minsupp }
 K=2
 While (Lk-1 ≠ ∅) do begin
 Ck = Apriori-gen (Lk-1) // Ck contains K-itemsets using Apriori-gen function
 For all transactions t belongs to Data Table do begin
 Ct = subset (Ck, t)//check weather Ck belongs to Transaction t
 For all candidates c belongs to Ct do begin
 c.wsupp+=hub(t)
 H+=hub(t)
 end
 Add Lk to result list
 Lk = { c belongs to Ck | c.wsupp/H > = minsup }
 end
 Increment k by 1
 End

5. Results

Table 1.First example of Transactions Table

TID	Item ID
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Table 2.The Data Table Formed

T id	I1	I2	I3	I4	I5
T100	1	1	0	0	1
T200	0	1	0	1	0
T300	0	1	1	0	0
T400	1	1	0	1	0
T500	1	0	1	0	0
T600	0	1	1	0	0
T700	1	0	1	0	0
T800	1	1	1	0	1
T900	1	1	1	0	0

Table 3.Results of Hub weights of Transaction

T id	Hub Weight
T100	148571
T200	80178
T300	122458
T400	139803
T500	116370
T600	122458
T700	116370
T800	205316
T900	182083

Table 4.Results of weighted supports of Items

Items	Weighted support
I1	0.736
I2	0.811
I3	0.701
I4	0.178
I5	0.287

When minimum support is set to average of weighted supports that is 0.452, output values of the frequent item sets are I1 I2, I1 I3, I2 I3 respectively.

Table 5.Results of Association Rules

Frequent Item sets	Strong Association rules	Confidence (%)
I1 I2	I1=>I2	47.58
I1 I3	I3=>I1	48.77
I2 I3	I3=>I2	46.36

When minimum support is decreased exponentially by power of 2 that is 0.226, output values of the frequent item sets are I1 I2 I3, I1 I2 I5 respectively.

Table 6.Results of Association Rules

Frequent Item sets	Strong Association rules	Confidence
I1 I2 I3	(I1 I3)=>I2	63.93
I1 I2 I5	(I1 I5)=>I2	55.78

Hence the Marketer can form various marketing strategies by keeping I1 I2 I3 and I1 I2 I5 together or I1 I2, I1 I3, I2 I3 together. Also it can be observed from the previous tables that item I2 (Maximum weighted support) is the most significant item and transaction T800 (Maximum hub weight).

Table 7.Second example of Transactions Table

TID	Item ID
100	A, B, C, D, E
200	C, F, G
300	A, B
400	A
500	C, F, G, H
600	A, G, H

Table 8.The Data Table Formed

T id	A	B	C	D	E	F	G	H
100	1	1	1	1	1	0	0	0
200	0	0	1	0	0	1	1	0
300	1	1	0	0	0	0	0	0
400	1	0	0	0	0	0	0	0
500	0	0	1	0	0	1	1	1
600	1	0	0	0	0	0	1	1

Table 9.Results of Hub weights of Transaction

T id	Hub Weight
100	17817737
200	14922313
300	8006294

400	5087331
500	18608778
600	14138943

Table 10.Results of weighted supports of Items

Items	Weighted support
A	0.573
B	0.329
C	0.653
D	0.227
E	0.227
F	0.426
G	0.607
H	0.417

When minimum support is set to average of weighted supports that is 0.384, output values of the frequent item sets is CFG.

Table 11.Results of Association Rules

Frequent Item sets	Strong Association rules	Confidence (%)
CFG	(F G)=>C	61.26

When minimum support is decreased exponentially by power of 2 that is 0.192, output values of the frequent item sets is ABCDE.

Table 12.Results of Association Rules

Frequent Item sets	Strong Association rules	Confidence (%)
ABCDE	(A B C D E)=>C	67.47

Hence the Marketer can form various marketing strategies by keeping ABCDE or CFG together.

Also it can be observed from the previous tables that C is the most significant item (Maximum weighted support) and

transaction 500 (Maximum hub weight) is the most significant Transaction.

6. COMPARISON OF ALGORITHMS

Here the Proposed Modified Apriori algorithm and Traditional Apriori Algorithm are compared on various points which clearly show how the proposed algorithm is better than the Traditional Apriori Algorithm.

Table 13. Comparison of Algorithms

Criteria	Modified Apriori Algorithm	Traditional Apriori Algorithm
Number of scans on Database	one	Many
Support	System generated	User supplied
Confidence	System generated	User supplied
Finding Significance of Transactions	Yes, Assigning Hub weights	No
Finding Significance of Items	Yes, Assigning Authority weights	No

7. CONCLUSION

Hence we have conceptualized and implemented the algorithm using .NET framework and SQL 2008 server. We have created a windows application software that can be installed by the company and perform Data Mining. The benefit of using this software is that the association rules are mined as per internal relationship between transactions and items and hence user input like minimum support and minimum confidence are not expected. Hence by using this software the company can enhance marketing strategies using weighted association rule mining.

8. REFERENCES

- [1] W. Wang, J. Yang, and P.S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," Proc. ACM SIGKDD '00, pp. 270-274, 2000.
- [2] Omiecinski, "Alternative Interest Measures for Mining Associations in Data- bases", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 1, pp. 57-69
- [3] K. Wang and M.-Y. Su, "Item Selection by "Hub-Authority" Profit Ranking, Proc. ACM SIGKDD, 2002.
- [4] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Datasets," Proc. ACM SIGMOD '93, pp. 207-216, 1993.
- [5] Ke Sun and Fengshan Bai "Mining Weighted Association Rules without Preassigned Weights", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 20, No. 4, pp. 489-495, April 2008.
- [6] J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," J. ACM, vol. 46, no. 5, pp. 604-632, 1999.
- [7] F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework," Proc. ACM SIGKDD '03, pp. 661-666, 2003.
- [8] Liu, B. Hsu, W., Ma, Y., "Mining Association Rules with Multiple Minimum Supports," Proc. Knowledge Discovery and Data Mining Conf., pp. 337-341, Aug. 1999.
- [9] A. Borodin, G. Roberts, J. Rosenthal, "Link Analysis Ranking : Algorithms, Theory, and Experiments", Panayiotis Tsaparas, ACM Transactions on Internet Technology, Vol. 5, No. 1, February 2005, Pages 231-297
- [10] R. Aggarwal, R.Srikant, "Fast Algorithm for mining Association Rules", IBM Almaden Research Centre, Proceedings of 20th VLDB Conference, Santiago, Chile, 1994.