ISSN 2047-3338

Algorithm for Weighted Association Rules for Sequential Data

N.K. Sharma[1]and Dr. R.C. Jain[2]
[1]Engineering College Ujjain, Ujjain, M.P., India
[2]S.A.T.I. Engineering College, Vidisha, M.P., India
[1]nksharma070965@gmail.com, [2]dr.jain.rc@gmail.com

***Abstract*– All the available algorithms working on arranging sequence in the database either in ascending or descending order. It is sometimes does not give result as per need. In the online admission process and in train / flight reservation, choices have significant role and if by using available algorithms for analysis purpose, the result will not be as per the requirement. So efforts have been made in this direction and design and implement an algorithm to overcome with such problem.**

***Index Terms*– Association Rule Mining, Database, Data Mining and Weighted Sequence**

I. INTRODUCTION

ACCESSING the data and storing that data for future reference is a common activity for researchers and business associates. Data mining is a valuable and interesting technique to retrieve valuable information from the stored data. The explosive growth in data collection in business and scientific fields has literally forced upon us the need to analyze and mine useful knowledge form it [1].

Data mining refers to the entire process of extracting useful and novel patterns from large datasets. Data mining tasks involve huge size of data and large amount of computation; hence high performance computing is an essential component for any successful large-scale data mining application. Data Mining and Knowledge Discovery in Databases (KDD) is a new interdisciplinary field merging ideas from statistics, machine learning, databases, and parallel and distributed computing. The key challenge in data mining is the extraction of knowledge and insight from massive databases [2], [3]. Association rule mining can be defined as a data mining method tries to distinguish interrelations of variables where large database exists.

Association rule mining techniques discover associations between itemsets, clustering techniques group the unlabeled data into clusters and classification techniques identify the different classes existing in categorical data. Most of the data mining techniques discover association rules from binary data while itemsets are mostly associated with quantity [7], [8].

II. WORK ALREADY DONE

C.H. Cai et al. [4] introduced the notion of weighted items to represent the significance of individual items. They also suggested that when we compute the weighted support of the rule, we can consider both the support and the important ratio (weights) factors. They also proposed two algorithms to mine weighted binary association rules. Abhinav Shrivastava et al. [5] proposed a new approach for database intrusion detection using data mining technique which takes the sensitivity of the attributes into consideration in the form of weights. The proposed approach mines dependency among attributes in a database. The transactions that do not support these dependencies are marked as malicious transactions. Yun Sing Koh and Gillian Dobbie [6] presented a approach that considers social network structure based on reach ability, and sociability of a researcher as a recommendations tool for potential collaborators.

In [7] D. Sujatha and Naveen CH investigated an approach to mine quantitative association mining, by decomposing categorical attributes into unique valued attributes and quantitative attributes into ranges. S.A. Sahaaya et al. [9] developed a weighted association rule mining algorithm to identify the best association rules that are useful for website restructuring and recommendation that reduces false visit and improve user navigation behaviour.

III. PROPOSED METHODOLOGY

The selected data must be preprocessed. In the preprocessing step data prepared for mining process. The output of the preprocessing step is abstract data. Than data mining algorithms applied to this abstract data. The preprocessing step contains another three separate phases. In First phase, the collected data must be cleaned, and removed

data such as name, contact number etc. In Second phase assigned encode to the user choice. In the third phase, raw data converted into a format suitable for the mining algorithm, duplicate, incomplete, noisy and inconsistent data items are eliminated in this phase:

- Retrieving counseling data
- Preprocessing
- Applying dynamic programming Technique
- Pattern discovery Phase
- Analyzing discovered Patterns

Let Sample Data have D have Transaction T = {T1,T2,T3,...,Tn} where Ti represent choice filled by ith student.

A set of available Choice C = {C1,C2,C3,...,Cm} and set of positive real number of weights W= {W1,W2,…,Wm} attached with each Choice.

Let's take a pattern of the form C1C3C2. In weighted representation, this pattern is represented as C1=3, C3=2 and C2=1. First Choice has the highest priority (P1=3), Second Choice hold highest priority less than one and last Choice contains the one (C2=1).

IV. PROPOSED ALGORITHM

Input:

D, a Database of Transactions
Nt , Total number of transactions in D
Ns, Number of subsets
P, Maximum available Choice

Output:

Find the Frequently selected choices

Method:

1. Scan the Database D, calculate the subset size Ds=Nt /Ns
2. Partition the dataset D into Ns number of subsets Ds1, Ds2,…, DsNs
3. Initialize the weight of each subset in D transactions into zero
4. for i = 1 to Ns do
5 for j = 1 to Ds do
6. If first Choice then
7. W=P
8. Else if next choice then
9. P=P-1
10. W=P
11 end if
12. Ds (j)=W
13. end for
14. end for
15. for i =1 to Ns do
16. R (i)=WFptree(Ds(i))
17. end for
18. Apply Dynamic programming technique to store R(i) into the OBST table.
19. Repeat
20. Scan the OBST data sets and call WFptree algorithm
21. Until reach the goal state
22. Obtain the best rule Rb
23. Return Rb

Algorithm Weighted FP Tree (WFptree):

Find frequently visited choices using weighted order representation.

Input:

Ts, a Transaction database Min-sup, the minimum support threshold value.

Output:

Fp, the frequently visited pages in D

1. Scan Ts and count the number of occurrences (No) of 1-Choice selection from user Choice record (Vp) using choice weights.
2. Compare No with minimum support count.
3. if No < min-sup then
4. Prune V p
5. End if
6. Sort frequent items in descending order based on their support.
7. Use this order when building the FP-Tree, so common prefixes can be shared.
8. FP-Growth reads 1 transaction at a time and maps it to a path.
9. Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix).
10. Find all frequent patterns containing one of the items.
11. Then find all frequent patterns containing the next item but NOT containing the previous one
12. Repeat (11) until we're out of items

Table I: Example (Sample Data)

Blocks	Name of Transaction	User Choice Order	Weighted order Representation			Binary Representation		
			C1	C2	C3	C1	C2	C3
B1	T1	C3-->C2-->C1	1	2	3	1	1	1
	T2	C1-->C2-->C3	3	2	1	1	1	1
	T3	C1-->C2	3	2	0	1	1	0
	T4	C1-->C2-->C3	3	2	1	1	1	1
	T5	C1-->C2-->C3	3	2	1	1	1	1
B2	T6	C2-->C3	0	3	2	0	1	1
	T7	C1-->C3	3	0	2	1	0	1
	T8	C1-->C2-->C3	3	2	1	1	1	1
	T9	C1-->C2	3	2	0	1	1	0
	T10	C2-->C1	2	3	0	1	1	0
B3	T11	C1-->C2-->C3	3	2	1	1	1	1
	T12	C3-->C2-->C1	1	2	3	1	1	1
	T13	C2-->C3	0	3	2	0	1	1
	T14	C3-->C2-->C1	1	2	3	1	1	1
	T15	C1-->C2-->C3	3	2	1	1	1	1
B4	T16	C1-->C2-->C3	3	2	1	1	1	1
	T17	C2-->C3	0	3	2	0	1	1
	T18	C1-->C2	3	2	0	1	1	0
	T19	C3-->C2-->C1	3	2	1	1	1	1
	T20	C1-->C2	3	2	0	1	1	0

Table II: Computational result of support

Blocks	P1->P2			P1->P2->P3		
	Support	Confidence	Lift	Support	Confidence	Lift
B1	4/5	4/5	1	3/5	3/5	1
B2	2/5	2/4	5/4	1/5	1/5	1
B3	2/5	2/4	5/4	2/5	2/5	1
B4	3/5	3/4	5/4	1/5	1/5	1

Blocks	P1-->P2			P1-->P2-->P3		
	Support	Confidence	Lift	Support	Confidence	Lift
B1 + B2	3/5	13/20	1	2/5	2/5	1
B2 + B3	2/5	1/2	5/4	3/10	3/10	1
B3 + B4	1/2	5/8	5/4	3/10	3/10	1

V. IMPLEMENTATION EXAMPLE DATABASE

The selected data must be preprocessed. In this preprocessing step data are prepared for mining process. The output of the preprocessing phase is abstract data. The data mining algorithms are applied to this abstract data. The preprocessing step contains three separate phases. First, the collected data must be cleaned. The third step, convert the raw data into a format needed by For example data such as name, contact number etc, are removed. Secondly, assign encoding to user choice. In the mining algorithm. Duplicate, incomplete, noisy and inconsistent data items are eliminated in this phase.

So the work is illustrated by considering 20 sample data from the entire data set (Table 1). The data is divided into four blocks B1, B2, B3 and B4. Every Block has five different transactions T1, T2, T3, T4 and T5 and so on. All the transactions are represented in weighted order and each block consist of three choice records with choice C1, C2 and C3.

VI. RULES GENERATED AND RESULT ANALYSIS

Table II shows the computational result of support, confidence and lift values for the two association rules such as C1-->C2 and C1-->C2-->C3 in B1, B2, B3 and B4. For instance, to the block B1 the Weighted order representation for the two Choice is C1=3 and C2=2 and its corresponding support and confidence values are 80% (=4/5) and 80% (=4/5). For the three Choice with weights such as C1=3, C2=2 and C3=1, its support and confidence values can be represented as 60% (=3/5) and 60% (=3/5). All the experiments are performed on a Xeon 2.8 GHz machine with 4 GB RAM running on Windows 7 platform. All the programs are written in JAVA platform.

VII. CONCLUSION

We have proposed to study a new approach of mining association rule using FP tree. In this approach, the choices are assigned weights to reflect their importance to the concerned users. The support, confidence, lift and number of rules have been significantly improved and better than the conventional association rule mining. The proposed method provides aspirants to assess his/her possibility of allotment of institutions of their choices on the basis of marks, caste, educational qualifications and location etc. the algorithm can be improved when approached with other techniques.

ACKNOWLEDGEMENTS

Our sincere thanks to our colleague Mr. Manoj Yadav, Software Consultant in Bhopal, Madhya Pradesh, India for the immense support you have provided us for publishing this paper.

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swamy, Database Mining: A Performance Perspective, IEEE Tran, in Proceeding of On Knowledge and Data Engg., December, 1991.

[2] M-S Chen, J Han and P. S. Yu, Data Mining : An Overview from a Database Perspective, IEEE Tran, in Proceeding of On Knowledge and Data Engg., December, 1996.

[3] R. Agrawal, T. Imielinski, and A. Swami, Mining Association Rules between Sets of Items in Large Databases, in Proceeding of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993.

[4] C.H. Cai, Ada W.C. Fu. C.H. Cheng and W.W. Kwong, Mining Association Rules with Weighted Items in Database Engineering and Applications Symposium, 1998. (Proceedings. IDEAS'98).

[5] Abhinav Srivastava, Shamik Sural and A.K. Majumdar, Database Intrusion Detection using Weighted Sequence Mining in Journal of Computers, July 2006

[6] Yun Sing Koh, Gillian Dobbie, Indirect Weighted Association Rules Mining for Academic Network Collaboration Recommendation in Proceedings of the Tenth Australasian Data Mining Conference (AusDM 2012), Sydney, Australia.

[7] D. Sujatha and Naveen CH, Quantitative Association Rule Mining on Weighted Transactional Data in International Journal of Information and Education Technology, August 2011

[8] Amir Hossein Azadnia, Shahrooz Taheri,Pezhman Ghadimi, Muhamad Zameri Mat Saman and Kuan Yew Wong, Order Batching in Warehouses by Minimizing Total Tardiness: A Hybrid Approach of Weighted Association Rule Mining and Genetic Algorithms in Hindawi Publishing Corporation The Scientific World Journal Volume 2013.

[9] S.A.Sahaaya, Arul Mary and M.Malarvizhi, Integrated Web Recommendation Model with Improved Weighted Association Rule Mining in International Journal of Data Mining & Knowledge Management Process (IJDKP), March 2013.