Review Paper: Procuring Frequent and Sequential Items to Improve Product Sales in E-Commerce Sites

Dr. M.Sreedevi
Department of Computer Science and Engineering
K L University
Guntur,India.
msreedevi_27@kluniversity.in

P. Haritha
Department of Computer Science and Engineering
K L University
Guntur, India.
harithapaladugu009@gmail.com

K. Ravali
Department of Computer Science and Engineering
K L University
Guntur, India.
ravali.kandregula@gmail.com

M. Manoj Pruthvi
Department of Computer Science and Engineering
K L University
Guntur, India.
manojpruthvim@gmail.com

Abstract-Increase in e-commerce industry has lead to availability of large amounts of data. Data is an important for everyone. There are hundreds of websites being deployed and each site offers millions of products. This means that there is a substantial amount of information being provided resulting in information overload and in turn results in reduced customer satisfaction and interest. Presenting frequent and sequential patterns in e-commerce sites results in increase of sales of products without delay. Different association rule mining techniques can be used for generating frequent and sequential patterns.

Keywords*-Frequent and Sequential pattern mining, Association rule mining, sequence rule mining, Horizontal database.*

*****

I. INTRODUCTION

Increase in e-commerce industry has lead to availability of large amounts of data. Data is an important for everyone. There are hundreds of websites being deployed and each site offers millions of products. This means that there is a substantial amount of information being provided resulting in information overload and in turn results in reduced customer satisfaction and interest. Presenting frequent and sequential patterns in e-commerce sites results in increase of sales of products without delay. Different association rule mining techniques can be used for generating frequent and sequential patterns.

II. LITERATURE SURVEY

Mining frequent and sequential patterns has an important role in wide e-commerce applications. Sequential pattern mining problem was first introduced by Agarwal and Srikanth [7] in, which discovers frequent subsequence as patterns in a sequence data base.

For mining sequential patters, different algorithms had been proposed (Generalized Sequential Pattern) algorithm is one such algorithms. GSP is based on Apriori-based approaches. It includes support for a pattern which is the number of sequences that contains the pattern [5]. When the support for a pattern is greater than the minimum support threshold, then this pattern becomes a frequent sequence. GSP mining method consider three nontrivial, inherent costs which are independent of detailed implementation techniques are:

1/Potential huge set of candidate sequences,
2/ Multiple scans of data bases
3/ Difficulties at mining long sequential patterns [8].

Another frequent sequence algorithm is Prefix Span which is a pattern growth method. Its main idea is to examine only the prefix sub sequences and project only their corresponding postfix subsequence into projected databases. Prefix Span approach mines complete sequential patterns faster than GSP. The disadvantages of this algorithm are, it doesn’t consider time constraints, time window and is not suitable for comparatively small databases.

Another algorithm which is an enhanced method of PrefixSpan, called EPSpan. First EPSpan find all largest-sequences which each data-sequence support. Then EPSpan prunes repetitions sequences that are generated from data-sequence. Consideration of time factors are not there in subsequent phase. At last all largest-sequences are passed to PrefixSpan as input Parameters for generation of sequential patterns.

Association rules are one of the data mining techniques which are used in generation of frequent patterns predominantly [7]. Association rules are the if/then statements that helps to discover uncover relationships between unrelated items in a database. Association rules are used to find the relationships between the products which are frequently used together. We have different association rulemining algorithms like AIS algorithm, SETM algorithm, Apriori algorithm, ApioriTID algorithm, AprioriHYBRID algorithm, FP-Growth algorithm. Some of the applications of association rules are catalogue design, market basket analysis, etc.Two basic criteria's that association rules used are support and confidence. Association rules are usually needed to satisfy a user-specified minimum support and a user specified confidence.MSreedevi et.al proposed closed regular pattern mining algorithms in different databases[2][3][6].

The rest of the paper illustrates different algorithms which are proposed recently for frequent sequential patterns and concludes the paper with references.

III. COMBINED ASSOCIATION RULE AND SEQUENCE RULE MINING ALGORITHM

For presenting frequent and sequential pattern items, the method that this algorithm used is, firstly for the given item set association rule mining technique is applied, by applying this we can able to obtain frequent product sets. And for these frequent item sets sequence rule mining technique is applied in order to know the sequence product sets.

A.Association Rule Mining

In e-commerce sites for improving product sales associated products plays a very important role. Association rule mining can be applied on data items to predict the frequent item sets. In this algorithm the frequent item sets are generated by scanning database several times. Support count of each individual item was accumulated during the pass over the database. Based on the minimum support count of these items the support count less than its minimum support count will gets eliminated from the list of the items. Candidate 2-item sets are going to be generated by extending frequent 1-items with other items in the transaction. In second pass over the database, support count of candidate 2-items are generated by scanning the given database and the count is checked against the minimum support threshold. Similarly those candidate (k+1)-item sets are generated by extending frequent k-item sets with items in the same transaction. Consider an example described below. Consider five transactions – I1,I2,I3,I4 and I5 and ten items i, ii, iii, iv, v, vi, vii, viii, ix , x and xi.

TABLE 1. SAMPLE TRANSACTION DATABASE

Transaction	**Items Purchased**
I1	{i, ii, iii, iv, v, vi}
I2	{vii, ii, iii, iv, v, vi}
I3	{i, viii, iv, v}
I4	{i, ix, x, iv, vi}
I5	{x, ii, ii, iv, v, xi}

Calculate frequent item set

Now, we assume that an item is said to be frequently bought if it is brought at least 60% of time So, we considerthe minimum support threshold value is 3. Now Table II, Table III, Table IV illustrates first level frequent items, second level frequent items and third level frequent items respectively.

TABLE II. FIRST LEVEL FREQUENT ITEM SETS

Item Sets	i	ii	iv	v	vi
Frequency	3	3	5	4	4

TABLE III. SECOND LEVEL FREQUENT ITEM SETS

Item Set	i, iv	ii, iv	ii, v	iv, v	iv, vi
Frequency	3	3	3	4	3

TABLE IV. THIRD LEVEL FREQUENT ITEMSET

Item Set	ii, iv, v	iv, v, vi
Frequency	3	2

Now from the consider example, we can conclude that {ii, iv, v} is the frequent set of items bought by the customers. The association of items need to find by using sequence rule mining.

B.Sequence Rule Mining

Sequence rule mining determines association between the products in a frequent item set. From the example we have frequent item set as {ii, iv, v}. Our aim is to find the association between ii, iv and v. If one product is purchased, we need to find the find probability that users will also buy the other products in the set.

In sequential rule, two measures are used: support and confidence. Support of a rule X→Y is the number of sequences containing item X followed by items from Y. Confidence of a rule X→Y is its support divided by the number of sequences containing the items from X. Table V gives sequence rules for the set {ii, iv, v} as follows:

TABLE V. SEQUENCE RULES

Sequence	Support	Confidence
{ii}→{iv, v}	3	3/3*100%=100%
{iv}→{ii, v}	3	3/5*100%=60%
{v}→{ii, iv}	3	3/4*100%=75%
{ii, iv}→{v}	3	3/3*100%=100%
{iv, v}→{ii}	3	3/4*100%=100%
{ii, v}→{iv}	3	3/3*100%=100%

If the confidence is 100%, then we can say that there are 100% chances that, if item set X are bought, then the products from set Y will also be bought.

IV. IMPROVED AC-APRIORI ALGORITHM

In the apriori algorithm, for the candidate k-sequence set C_k, when calculating its support, it is necessary to traverse the data base all the time. Here repeated scan is happening which leads to decease in efficiency of apriori algorithm.

We have the cases where items bought in one transaction may be repeated within the same transaction. At that time, scanning of database all the time is not an efficient process. To work at that case also a new algorithm called AC-Apriori algorithm is used. It implies Aho-Corasick automation. In this algorithm the method that is followed is, while calculating the support for $C_{k.,}$ the Trie tree and the failure pointer are constructed for C_k and C_k is converted to AC automation AC-C_k. For each transaction, we only need to search AC-C_k once, we can know which k-sequence in C_k is included in the transaction..we considered here a sample sequence items list purchased

AC-Apriori Example: minimum support is 3.

TABLE VI. SAMPLE SEQUENCE DATABASE

Transaction ID	Sequence Items Bought
I1	iii, i, ii, iii
I2	i, ii, iii, ii, v
I3	iii, i, ii, iii, iv
I4	i, ii, iii, i
I5	iii, iv, ii, v

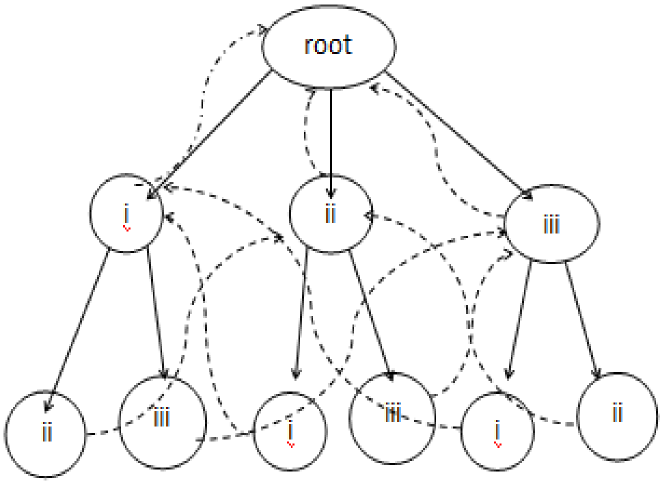

Figure 1. AC automation AC-$C_{2.}$

Frequent 1- sequence set:

L1= {{{i},{ii},{iii}}

TABLE VII. SUPPORT OF CANDIDATE 2 - SEQUENCESET

Candidate 2-Sequence	Support
i, ii	4
i, iii	0
ii, i	0
ii, iii	4
iii, i	3
iii, ii	1

Here, the support is calculated by considering the sequence which has obtained from the AC-C_2 tree. The sequence which has obtained from tree is checked against the data base and accordingly for the sequence thecount gets incremented. By comparing with minimum support i.e. 3, the frequent 2- sequence set is given as:

L2={{i, ii},{ii, iii},{iii,i}}

Now the Candidate3-sequence

set:C3 ={{i, ii, iii},{ii, iii, i},{iii, i, ii}}

C3 is converted to AC automation AC-C3 in figure 2.

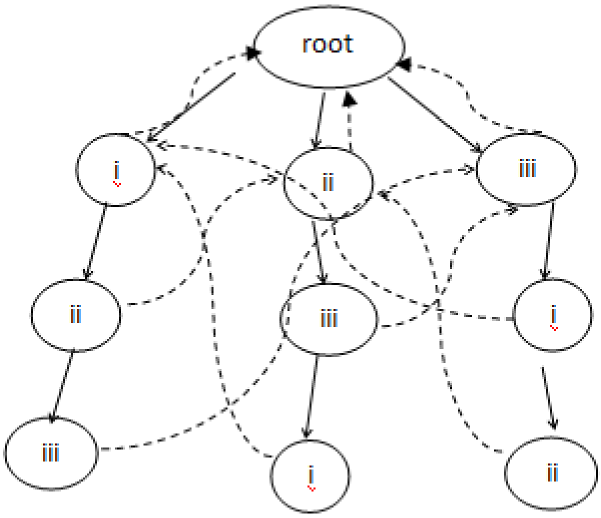

Figure 2. AC automation AC-C_3

TABLE VIII. SUPPORT OF CANDIDATE 3 - SEQUENCE SET

Candidate 3-Sequence	Support
i, ii, iii	4
ii, iii, i	1
iii, i, ii	2

Frequent 3-sequence set:

L3= {{i, ii, iii}}

Since, we have obtained only one frequent sequence set, there is no need of going to further candidate generations.

Now, we can say that the sequential patterns are: {{i},{ii},{iii},{i, ii},{ii, iii},{iii, i},{i, ii, iii}}.

V. V.HORIZONTAL DATABASE FORMAT ALGORITHM

This algorithm is based on apriority algorithm, which means this algorithm have properties to discover intra transaction association and by using this method we can generate rules to discover associations.

Generally, first version of horizontal database is considered in 5 steps, they are

a) Sort phase
b) Large Item Set
c) Transformation Phase
d) Sequence phase
e) Maximal phase

For clear understanding of horizontal database format, let us consider a (TABLE IX) sample customer transaction with sample data which consists of customer id, transaction time and item bought.

a.Sort Phase

In sort phase, it sorts the data from original table (TABLE IX) sample customer transaction to (TABLE A1) customer transaction database by customer id and transaction time.

b. Large Item Set Phase

In large item set phase, it finds out all the set of large item sets, these large item sets need to meet minimum support, for example minimum support of 25%[9].

c. Transformation Phase

In Transformation phase, each customer sequence is changed by replacing each transaction with set of items set in the transaction. Transactions which are not having any large item set are not contained to hold and a customer sequence which are not holding any large item sets are eliminated.

d. Sequence Phase

In sequence phase data will be mined for frequent subsequence's. The process begins with the largest sequences and terminates when either no elements are generated or no element meets the minimum support criteria.

e. Maximal Phase

In Maximal Phase we will find all maximal sequences among the set.

TABLE IX. SAMPLE CUSTOMER TRANSACTION

Customer Id	Transaction Time	Item Bought
1	25/10/17	iii
1	30/10/17	viii
2	10/10/17	i, ii
2	15/10/17	iii
2	30/10/17	iv, vi, vii
3	15/10/17	iii, v, vii
4	25/10/17	iii
4	15/10/17	iv, vii
4	20/10/17	viii
5	30/10/17	viii

TABLE A 1 CUSTOMER TRANSACTION DATABASE

Customer Id	Customer Sequence
1	(iii) (viii)
2	(i ii) (iii) (iv vi vii)
3	(iii v vii)
4	(iii) (iv vii)(viii)
5	(viii)

VI. CONCLUSION

There are various algorithms for mining frequent and sequential patterns and in this paper and in this we have presented two different algorithms named combined association and sequence rule mining algorithm, AC apriori algorithm and Horizontal database format algorithm Each algorithm has its own significance. According to input considered, we can use any of these algorithms for knowing frequent and sequential items.

REFERENCES

[1] Z.A.Usmani,ShraddhaManchekar, TahreemMalim, Ayman Mir, “A Predictive Approach for Improving the sales of Products in E-commerce”, 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics,2017.

[2] M Sreedevi and L.S.S.Reddy” Mining Regular closed in transactional databases”, IEEE Conference 2012 page No 380-383

[3] M Sreedevi and L.S.S Reddy “Parallel and Distributed closed regular pattern mining in large databases” IJSCI.org, Volume 10 Issue 2 No 2 March 2013 Page No 264-269

[4] Jun Yang, Haoxiang Huang, XiaohuiJin,”Mining Web Access Sequence With Improved Apriori Algorithm”, IEEE International Conference of Computational Science and Engineering(CSE),2017.

[5] YemingTang,Quili Tong, Zhao Du “ Mining frequent sequential patterns and association rules on campus map system”, 2nd International Conference on Systems and Informatics,2014.

[6] M.Sreedevi and L.S.S.Reddy ”Closed Regular Pattern Mining using Vertical Format” IJSCET ,Volume 4 ,No 7 July 2013 Page no 1051-1056

[7] Trupti A. Kumbhare, Santosh V.Chobe, ”An Overview of Association Rule mining Algorithms”, International journal of computer science and information technologies,2014

[8] Jia-Dong Ren, Yin-Bo Cheng, Liang-Liang Yang,” An Algorithm for Mining Generalized Sequential Patterns”, Proceedings of Third International Conference on Machine Learning and Cybernetics,2004.

[9] Mooney, C. H. and Roddick, J. F. 2013. Sequential pattern mining – Approaches and algorithms. ACM Comput. Surv. 45, 2, Article 19 (February 2013), 39 pages.