International Journal of Engineering & Technology, 7 (1.1) (2018) 273-277

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A survey for acquiring frequent and sequential items in E-commerce sites

Haritha P *, Sree Devi M, Ravali K, Manoj Pruthvi M

Koneru Lakshmaiah Educational Foundation, Andhra Pradesh, India
**Corresponding author E-mail: harithapaladugu009@gmail.com*

Abstract

Large amounts of data has made available because of the increase in e-commerce industry. Data has high significance and also important for everyone. Hundreds of websites are being deployed and each site offers millions of products. In addition to this there are several types of input forms. Different sites have different input item collection. 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. This huge amount of data needs to get processed so that we can able to extract the useful information. From this useful information we can able to increase customer interest, satisfaction along with sales of e-commerce sites. Presenting frequent and sequential patterns in e-commerce sites results in increase of sales of products without delay. Different association rule mining techniques and sequential rule mining techniques can be used for different sets of input forms in order to generate frequent and sequential patterns. This paper discusses various algorithms using techniques such as association rule mining, sequence rule mining proposed for mining frequent and sequential items.

Keywords: *Data Mining; Frequent Pattern Mining; Sequential Pattern Mining; Association Rule Mining; Sequence Rule Mining*.

1. Introduction

Data mining is an activity of examining the data from divergent views and outlining it in order to produce valuable information [7]. Now-a-days the predominant usage of Internet has led to the rise of e-commerce industry, the consumer driven industry. Because of the availability of hundreds of e-commerce websites and each site offering millions of products, it has created difficulty for the users to browse through the products and make decisions [1]. It is also becoming complex for e-commerce administrators also to have effective decisions which leads to improve the market. Enhancing the market of every e-commerce website is becoming complex due to millions of products. Data mining is an analysing tool that is used for analysing the data. It allows users to categorize the give data and present relationships among data. Different techniques are used for data mining like association techniques, classification techniques, clustering techniques and soon. Data mining is the process of sorting which is done on large data sets. By applying data mining on data sets, we can able to identify several useful patterns and also we can develop relationships which can be used to solve problems through data analysis. Data mining is widely used in enterprises to predict future trends.

2. Literature review

Mining frequent and sequential patterns has an important role in wide e-commerce applications. Sequential pattern mining complication was first intiatiated by Agarwal and Srikanth [7], which detects frequent subsequence as patterns in a sequential data base.
For mining sequential patters, different algorithms had been proposed Generalized Sequential Pattern algorithm is one such algorithm. Other algorithms we have are Prefix Span algorithm and also EPSpan algorithm.

2.1. Generalized sequential pattern algorithm

GSP is based on Apriori-based approaches. It includes support for a pattern which represents the sequences that includes required pattern [5]. When the support for a pattern is exceeding the minimum support threshold, then this pattern becomes a frequent sequence. GSP mining method includes three nontrivial and also cost inherent that are independent of comprehensive implementation techniques. They are:

- Candidate sequences of potentially huge sets
- Databases including several scans
- Complications involved in mining long sequential patterns [8].

2.2. Prefix span algorithm

Another frequent sequence algorithm is Prefix Span. It is a Pattern growth method. Main proposal is to investigate prefix sub sequences only and assign only their corresponding postfix subsequence into projected databases. A projected database includes sub sequences sets residing in database. The subsequence represents suffixes of sequences that have prefix. At every step, the algorithm focuses on frequent sequences that includes prefix in corresponding database. By this approach, the search space is reduced at each step which results in allowing for better performances in the presence of small support thresholds too. Prefixspan algorithm uses the prefix projection technology which could eliminate candidate items effectively [9].
Prefix Span approach mines complete sequential patterns faster than GSP. The disadvantages of this algorithm are, it doesn't con-

sider time constraints, time window and is not suitable for comparatively small databases.

2.3. EP span

Another algorithm which is an enhanced method of PrefixSpan, called EPSpan. First EPSpan notices all largest-sequences. They can be found by data-sequence support. Then EPSpan trims repetition sequences that are produced from data sequence. Considerations of time factors are not there in subsequent phase. At last all resulted largest-sequences are given as input parameters to PrefixSpan in order to generate sequential patterns.

2.4. Association rule mining

Association rules are one of the data mining techniques which are used in generation of frequent patterns predominantly [7]. Association rules are if/then statements. Within a database, they assist to discover uncover relationships among unrelated items. Products which are frequently used together can be known by association rules. We have different association rule mining 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. There are two fundamental benchmarks that association rules employed are support and confidence. Rules of association are conventionally needed to fulfill minimum support and confidence specified by user. M Sreedevi 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.

3. Consolidated 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 [1].

3.1. 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, item sets which are frequent are generated by inspecting the database multiple times. Each individual’s items support count was accumulated during the pass over the database. Based on the minimum support count of items, items for which support count is less than its minimum support count will be discarded from the list of the items. Hence, Candidate 1-item sets are generated by considering frequent 1-items. Now, candidate 2-item sets are generated by extending frequent 1-items with other remaining items in the transaction set. During second pass on database, support count for candidate 2-items are generated by scanning the given database and the count is verified against minimum support threshold. Likewise the candidate (k+1)-item sets are generated by including frequent k-item sets with the items within transaction. Consider an example described below. Consider five transactions – T1, T2, T3, T4 and T5 and ten items I, II, III, IV, V, VI, VII, VIII, IX, X and XI.

Table 1: Sample Transaction Database

Transaction	Items Purchased
T1	{I, II, III, IV, V, VI}
T2	{VII, II, III, IV, V, VI}
T3	{I, VIII, IV, V}
T4	{I, IX, X, IV, VI}
T5	{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 consider the minimum support threshold value is 3. Now Table 2, Table 3, Table 4 illustrates first level frequent items, second level frequent items and third level frequent items respectively.

Table 2: First Level Frequent Item Sets

Item Sets	I	II	IV	V	VI
Frequency	3	3	5	4	4

Table 3: Second Level Frequent Item Sets

Item Set	I, IV	II, IV	II, V	IV, V	IV, VI
Frequency	3	3	3	4	3

Table 4: Third Level Frequent Item Sets

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 needs to find by using sequence rule mining.

3.2. 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 5: 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. From the sample example, we can come to a conclusion that if any person buys item < II > then there is a chance of buying < IV > and < V > items to a maximum extent. We can also say that if a person buys < IV > then the chances of buying the items < II > and < V > is very less. Hence we can use this algorithm for representing frequent and sequential items for the similar sample input forms.

4. Improved prefix span algorithm

For generating frequent and sequential patterns we have studied another kind of algorithm which is named as BLSPM. For sequential pattern mining we have a lot of classic algorithms. Among those algorithms, Prefix span algorithm is one of the most widely used algorithms. Prefix span algorithm make the usage of prefix projection technology [9], which could avoid candidate items effectively, thus to improve the mining efficiency in a certain extent. Prefix span algorithm includes the structure with lot of projection database. Here structuring projection database consumes huge memory. It also increases the scan time. Hence we have considered an algorithm named BLSPM which is an enhanced way of prefix span algorithm. The major difference between BLSPM and prefix span algorithm is BLSPM algorithm make effective branches reduction.

The algorithm includes a parameter named weighted value of item. Weight value is a sequence of entries in the database for each non-negative real number. To recognize the importance of each item, we use the entry of weight value. Weight value make free of data units of restriction, which was converted into dimensionless pure value in favor of different order of indicator or magnitude units which can be compared or weighted. Min-Max standardization is a linear transformation of original data in which results are mapped to between [0-1]. Conversion functions are as follows: P* = (P – min)/ (max-min) where P* represents transformed value of P, original data represents the minimum value min; max indicates the maximum value of sample data. If P represent a frequent sequence, P = < S1 S2 …S1> frequent item set P weight value equal to the frequent item sets of weight accumulation and frequent item sets length. i.e.., Weight (P) = ∑Weight (P)/ length (P).Here ∑ will go up to length (P). Weight (P) represents the weight of heavy frequent sequence P. The algorithm includes another parameter named sequence mode value VSP (P). Sequence mode value is frequently used to measure the importance of sequence, the main reference weights and branched. It is equal to weights * support. VSP (P) = Weight (P) * Support (P). BLSPM represents a barrier projection and pruning strategy combined with each other in a way that it can greatly decrease the number of scanning projection database, which results in improved efficiency. BLSPM include the following procedure (1) Make an effective reduction branches, In building a projection database, discard the items that has support value less than the required minimum support. We can treat them as infrequent items and thus we can remove from sequence database. (2) Applying the method of bi-level projection i.e.., if the sequence length is odd, we can construct the original projection database, if the sequence length is even, instead of constructing a projection database, we can construct a lower triangular matrix which decreases the time of scanning projection databases. (3) Now we can reorder the results of the sequences which are obtained based on the size of the sequential pattern values. By this it has made possible to find the most important sequence patterns.

Consider the sample transaction database with includes four transactions and items bought. Here we can consider sample items as I, II, III, IV, V, VI, and VII. Consider support value i.e.., min_sup = 2 for the sample example.

Table 6: Sample Transaction Database

Transaction ID	Sequence Items
T1	<I(I,II,III)(I,III)IV(III,VI)>
T2	<(I,IV)III(II,III)(I,V)>
T3	<V(I,II)(IV,VI)III,II>
T4	<V,VII(I,VI)III,II,III>

Now for the given sample sequence items, calculate the first scan sequence database to find all frequent sequences of length 1.

Table 7: Frequent Sequences of Length 1

Items	Frequency
I	4
II	4
III	4
IV	3
V	3
VI	3
VII	1

Since <VII> item support is less than the minimum support, min_sup. So, <VII> is non-frequent items and it needs pruning rounding.

Since the sequence length which is 6 is even, we can construct a lower triangular matrix. The lower triangular matrix is constructed by considering the length of frequent item 1 as X<Y axis. It is constructed based on weighted value of item. Construct a 6x6 lower triangular matrix as shown below:

Table 8: Matrix S

I	2					
II	(4, 2, 2)	1				
III	(4, 2, 1)	(3, 3, 2)	3			
IV	(2, 1, 1)	(2, 2, 0)	(1, 3, 0)	0		
V	(1, 2, 1)	(1, 2, 0)	(1, 2, 0)	(1, 1, 0)	0	
VI	(2, 1, 1)	(2, 2, 0)	(2, 2, 0)	(1, 1, 1)	(2, 0, 1)	1
	I	II	III	IV	V	VI

In the table matrix S is a matrix that shows all sequences of length 2 mode support. In the matrix there is a diagonal cell count, M [I, I] = 2 which represents that sequence< I, I > appears twice in S matrix. Other cells which are having corresponding three counts, such as M [I, II] = (4, 2, 2) represents that < I, II > is of support 4, < II, I > support degree is 2. Table S is a lower triangular matrix, which is based on both sides of the diagonal line because the position is symmetrical and for symmetrical position the count is same. Now for all the frequent items mined, calculate the value of each frequent item sequence pattern set. According with the sequence mode value descending order of items is arranged. In the way to arrangement if the sequence mode of any two frequent tem sets has the same value, then they are arranged according to alphabetical order. Consider the weight value for standardized sequence database of each item after distribution.

Table 9: Sequence Database Heavy (Weight) Table

Item	Normalized Weight
I	0.8
II	0.3
III	0.7
IV	0.5
V	0.8
VI	0.1

For example consider < III > prefix, Sequence mode value, VSP (III) = 1.4, VSP (III, I) = 1.5, VSP (III, II) = 1.5, VSP (III, III) = 2.1. According to results items are sorted as {< III, III >, < III, I >, < III, II >, < III >}, the process can be repeated for all the other items. The final results are shown in the table.

Table 10: Final Frequent Sequential Pattern

Prefix	Frequent Sequential Pattern
< I >	< I, III >, < I, II >, < (I, II) >, <I, III,III >, < I, III, II >, < I, I >, < I, III, I >, < I, IV, III >, < I, IV >, < I, (II, III), I >, <I, II, I >, < I, II, III >, < I, (II, III) >, < (I, II), III>, < (I, II), IV, III>, < (I, II), IV >, <I, VI >, < I >, < (I, II), VI >
< II >	< II, III >, < II >, < (II, III), I >, < II, I>, < (II, III) , < II,VI >
< III >	< III, III >, < III, I >, < III, II >, < III >
< IV >	< IV, III >, < III >, < III, III >, < III, I >
< V >	< V >, < V, I >, < V, I, III >, < V, III >, < V, I, III, II >, < V, I, III >, < V, II, III >,< V, III, II >, < V, II >, < V, VI, III >, < V, VI, III, II >, <V, VI >, < V, VI, II >
< VI >	< VI, III >, < VI, II, III >, < V, III,II >, < V, II >, < V >

By using BLSPM which is an improved method of prefix span algorithm, we can able to generate frequent and sequential patterns. This algorithm can be used for the sequence input forms.

5. 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 [4]. 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 11: Sample Sequence Database

Transaction ID	Sequence Items Bought
T1	III, I, II, III
T2	I, II, III, II, V
T3	III, I, II, III, IV
T4	I, II, III, I
T5	III, IV, II, V

Frequent 1- sequence set:
L1= {{{I}, {II}, {III}}

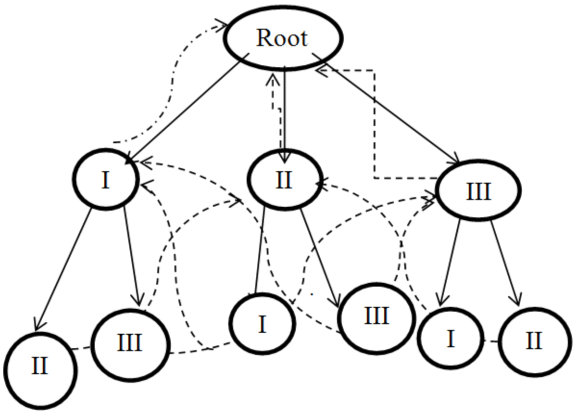

Fig. 1: AC Automation AC-C2.

Table 12: Support of Candidate 2-Sequence set

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 the count 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.

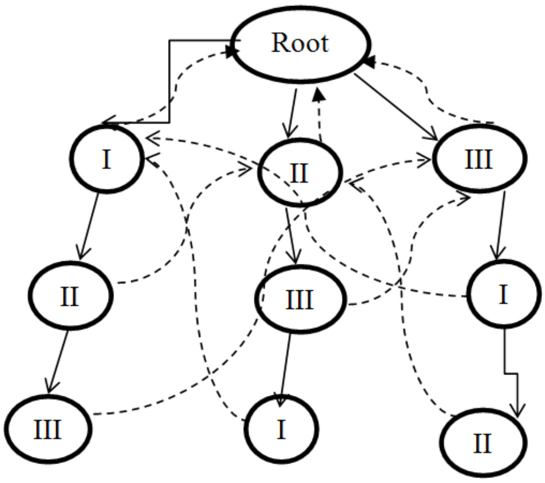

Fig. 2: AC Automation AC-C3.

Table 13: 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}}.

6. Horizontal database format algorithm

This algorithm is based on apriority algorithm, which means these algorithms 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 14) 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 14) sample customer transaction to (Table 15) customer transaction database by considering parameters 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% [15].

c) Transformation phase
In Transformation phase, for each customer, the sequence is changed by substituting each transaction with set of items 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 subsequences. The process starts with largest sequences and terminates when there are no elements to be generated or no element reaches to minimum support criteria.

e) Maximal phase

In Maximal Phase, all maximal sequences among the set are retrieved.

Table 14: 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 15: 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)

7. Conclusion

There are various algorithms that can be used for mining frequent and sequential patterns and in this we have presented different algorithms named consolidated association and sequence rule mining algorithm, BLSPM 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,Shraddha Manchekar, Tahreem Malim, 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, Xiaohui Jin,”Mining Web Access Sequence with Improved Apriori Algorithm”, IEEE International Conference of Computational Science and Engineering (CSE), 2017.

[5] Yeming Tang,Quili Tong, Zhao Du “ Mining frequent sequen-tial 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 Algo-rithm for Mining Generalized Sequential Patterns”, Proceedings of Third International Conference on Machine Learning and Cybernetics,2004.

[9] Peng Huang,” Improved algorithm based on Sequential Pattern Mining of Big Data Set “, IEEE, 2016. https://doi.org/10.1109/ICSESS.2016.7883028.

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