RESEARCHERID
THOMSON REUTERS

[Elakkiya, 6(9): September 2019] ISSN 2348 – 8034
DOI- 10.5281/zenodo.3461151 Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

MAX-CLOSED SPAMBY USING DIRECT BIT POSITION METHOD

Dr. E. Elakkiya*[1] & Dr. S. Ravichandran[2]

[*1]Teaching Assistant, Department of Computer Application, Alagappa University, Karaikudi – 630001, Tamil Nadu, India,

[2]Head of the Department, Research Department of Computer Application, H.H. The Rajah's College, Pudukkottai -622004,Tamil Nadu, India

ABSTRACT

This paper is Max-Closed SPAM by using Direct Bit Position (Maximal Closed sequential pattern mining). This algorithm is to acquire the maximal Closed Sequential Pattern from Sequence Database. The sequence pattern mining gives more numbers of patterns, closed sequential pattern obtains little number of patterns and maximal closed attain very few sequence patterns. According to the closed Sequential Patterns is long sequence database, the maximal closed pattern is influential by memory and performance. Experimental evaluation has done on UCI Repository Datasets, that shows the algorithm Max Closed DBP SPAM is efficient than the previous Closed DBP algorithms. The maximal closed frequent patterns are retrieved efficiently by given specified minimum support threshold value.

Keywords*: Pattern Mining, Binary Representation, Maximal Closed Pattern, Closed Sequential pattern*.

I. INTRODUCTION

The entire world has enormous information even over the atmosphere of the earth. Both living and non-living things have certain data to handle every day for their survival. The Sequential pattern mining technique is introduced by Agarwal[1].The main role of sequential pattern mining plays significant and it is important to a wide range of applications, such as the analysis of web click-streams, program executions, medical data, biological data and e-learning data [2].These types of method are significant to take a superior choice for better business solutions. Still researchers are finding novel patterns mining algorithm for a large sequence databases in many different ways. Plenty of algorithms are created and used for Data mining in the world. However, the problem is to generate candidates on mining sequential patterns in enormous sequence database and execution time as well. Discovering the all maximal closed frequent sequential patterns are challenging as the search space is tremendously large.

II. PROBLEM DEFINITION

In SPAM, maximum size of patterns is less but it contains many closed patterns within. Initially, closed frequent sequential pattern mining was introduced by Pasquier et al. in ICDT'99[3].This pattern is redundant because its supports can easily derive by its super-patterns with the same supports. The closed frequent sequential patterns are frequent patterns it doesn't have frequent super-pattern with the same support threshold value. In this method, no generating those redundant patterns; mining procedure is able to be more efficient. Basically Colspan incorporates some pruning techniques into PrefixSpan to find the closed set of frequent sequential patterns.

Though most of the previous method tackles the two factors in a certain degree, the property of item ordering in a sequence are not fully utilized in the mining process. Therefore, in this paper, proposed method called DBP-Maximal Closed SPAM for mining the maximal closed sequential patterns from frequent patterns. Still some pruning methods are used to support, repression and positional data.

III. MINING BASICS

Let A={x1, x2…., xm} is a set of items, an itemset is a subset of B. A sequence S ={s1, s2…,sm} is n order list of item sets. Thelength of S is m, which is the number of item sets, and S is also known asm-Sequences. A Sequence A=(x1, x2,...,xm) is a sub-sequence of another sequence B=(y1,y2,...,yn) if there exists a set of indices m1,m2,..,mi,n<=m1<=m2<=...<mi<=j. such that x1<=ym1,x2<=ym2,..., xi<=ymi. Then again, B is called a super-sequence of A. we can say that Bincludes A, or Aisenclosed by B. A sequence database D holds a set of sequences, and the support of a sequence S is the number of sequences that contain S. A frequent sequence is a sequence with support not less than the minimum support threshold, *min_sup*. A closed frequent sequential pattern is a frequent sequence that doesn't have any frequent super-sequence with the same support threshold value. *A maximal closed pattern is a longest sequence that does not have any frequent super sequence in Database(patterns that are not included in another pattern).*

IV. PROPOSED APPROACH

In this approach makes an item bit position table for all the sequences in the sample database D. Consider a sequence S1=<a (cd) ad>. The items position are found travel around the sequence left to right and its corresponding positions are stored. The length of the binary represented row in the position table is equal to the length of the sequence in the database D. When the item X is in the i^{th} position of the sequence from left, the i^{th} position of that item X is placed to 1, otherwise it is placed to 0.

In the first stage, it scans the sequence database at one time to record the positional information of each different item set in the database. Then it can simply gain all the frequent item sets that is length 1-sequences by accomplish their positional information. The positional data of an item i, represented by POSi, it consists of a lot of pairs of (sid,eid), where sid is the sequence identifier and eid is the element identifier. Because sid points out that sequence item lies in and eid indicates in which order the item lies in the sequence, this representation can reserve the information of item ordering without any loss. Let us assume a sequence database in tabel1.

Table1: Sample Sequence Database D

Sid	Sequence
S1	<a(c d)a d>
S2	<a c a e>
S3	<c a d (b c d)>
S4	<b b c>
S5	<(b c d)d>

The minimum support threshold value=2, the positional data of items as shown in table2.

Table 2: Positional Data for a Sequence

Sequence < a(c d)a d>				
S1	**a**	**cd**	**a**	**d**
a	1	0	1	0
c	0	1	0	0
d	0	1	0	1

There is no difficulty to managing lexicographical prefix tree in this method. Hence, this is an efficient than the previously methods for closed SPAM.

Consider the sequence S1 in the sample database D, there are four elements and four sequences as shown in the table2. If the item is present in the sequence, it is represented by 1 otherwise denoted by 0. It shows clearly on the table2. Since ‘a’ is present in the sequence 1 and 3, the bit position corresponding to ‘a’ is 1010.All the sequence constructs the same way for positional data.

To decrease the computational cost of checking bits in the position table, Item presence table is constructed with three fields namely Item, Sid and supports (min_sup.). Here again use top down approach and recorded the item present in the sequence of database S. If an item is present in i^{th} row of the sequence database, then it is assigned by 1, else it is assigned by 0. Consider the item “a” in sample sequence database S, since “a” is present in S1, S2, S3 the item presence table is constructed as I = 1, 1, 1, 0, 0. The complete item Present In table is constructed like shown in the table3.

Table 3: PresentIn Table

Item	S1	S2	S3	S4	S5	Sup.
a	1	1	1	0	0	3
b	0	0	1	1	1	3
c	1	1	1	1	1	5
d	1	0	1	0	1	3
e	0	1	0	0	0	1

The candidates are straight from the position and present. In these shown tables are explained without any doubt. As an alternative of generating the candidates by inserting a data into pre-known frequent patterns, the proposed approach directly generates the candidates by using the bit position table and the Present In table. Let us consider that the given *min_sup* threshold value given by the user is 2. From the Item presence table, the item “e” is pruned since the support is 1, it is less than the *min_sup*=2. (i.e.) sup (e)<*min_sup*. Consider the presence table to create candidates, firstly, the item ‘a’ and ‘b’ measured. a = {1 1 1 00}, b = {0 0 1 1 1}, at this time to find (a)(b) and (“a” “b”) The operation AND is performed in the Presence table values of ‘a’ and ‘b’ like, the previous DBP-SPAM.

Example1: Let’s take table1 is the input sequence database D. If the min_sup=2, the Maximal closed sequential is MCS={(aa):2, (aca):2, a(cd):2, (add):2, (bcd):2, (bd):2, (cad):2, (cd)d:2} from the FS have 13 sets of sequences ={(aa):2, (ab):3, (aba):2, a(cd):2, (ad):2, (add):2, (bc):2, (bcd):2, (bd):2, (ca):3, (cad):2, (cd):3, (cd)d :2}

Example2: Let’s have the table4 is a sample sequence database, referred as D1 when the perspective is unambiguous. The alphabetic order is taken as default lexicographical order. if min_sup=2,Maximal Closed Frequent Patterns (MCS)={(a f)d: 2, (e a b): 2}, Closed Frequent Sequential Pattern (CS) ={(af)d:2, (ea):3, (eab):2} while the corresponding Frequent Sequence (FS) set has 16 sequences. CS has the exact same information as FS, but includes much less frequent patterns.

Table 4: Sample Database

SID	Sequence
1	<(af)dea>
2	<eab>
3	<e(abf)(bde)>

V. ALGORITHM OFMAXIMAL CLOSED DBP-SPAM

In step1, the Sequence database is scanned to the Presence Table for 1-sequence. Step2, begin with 2-sequenceAND operation with the generated candidates at the same time, pruned the items, if items have less than the *min_sup* value. Step3 is to generate the I-extended and S-extended frequent items with its supports. In step4, the (MCS) Maximal Closed Frequent Sequences are retrieved from the complete (CS) Closed Frequent Sequence by comparing the each FS item sets.

[Elakkiya, 6(9): September 2019] ISSN 2348 – 8034
DOI- 10.5281/zenodo.3461151 Impact Factor- 5.070

VI. PSEUDO CODE FOR MAXIMAL CLOSED DBP-SPAM

Maximal dbp closed spam (S, min_sup)
INPUT: S–Sequence Database, min_sup-Minimum Support,
OUTPUT: Max-Closed Sequential patterns
INPUT: Sequential database D, min-sup
OUTPUT: Max-Closed Sequential patterns
BEGIN:
For each [Sidi, S]<= D begin
For each Element Sj of s begin
For each item i<=Sj begin
If Present In(i) = 0, Mark Present In(i) = 1
Set j^{th}bit in POSI(i) =1
End for
End For
End For
Patterns= IS_patterns(Present In, min-sup)
Closed Patterns=Max Closed DBP_SPAM(Patterns,min-sup)
END

Function Max Closed DBP_SPAM(Patterns, min-sup)
INPUT: Sequential Patterns, min-sup
BEGIN:
For each Patterns(i)< Patterns(n) begin
For each Patterns(j=i+1)< Patterns(n) begin
//check the pattern is Maximal closed or not
 //check pattern(i) is super or Sub sequence of Pattern(j)
If Patterns(i) is like Patterns(j) then
Next
Else
Return Patterns(i)
End if
Else
Return Patterns(i)
End if
End For
End For
END

Function IS_Pattern(Present In, min-sup)
INPUT: Present In table, min-sup
BEGIN:
For each item i <= Present In Table
Form base item sets by applying AND operation
If base Itemsets>= min-sup
Store base item sets in base table
End for
For each Itemsets K <= Base table
Fetch POS tables according to the items <= K
Find S_Extended patterns based on position
Count the S_extended patterns
If S_extended patterns >= min-sup

Store in Results
Find I_Extended patterns based on equal position
Count I_extended patterns
If I_Extended patterns >=min-sup
Store in Results
End For
Return Results

VII. EXPERIMENTAL EVALUATION

The proposed Maximal Closed DBP-SPAM is implemented on Visual C# programming on a personal computer of Intel 2.66 GHz Dual Core processors, 2GB RAM on Windows7, 32bit Ultimate. The experimental evaluation is performed on real world UCI Repository data sets. It is a transactional data set which contains all the transactions taken place between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. It is downloaded from the internet. The sample real dataset of UCI Online Retail is shown on table5 and the characteristics in table 6.

The figure1 shows, the performance analysis of proposed algorithm, the experimental evaluation is concerning with the running time is compared to UCI real world Online Retail dataset. The result as the minimum support value is changed from 0.01 to 0.05 percentages. The experiments are carried out with varying *min_sup* values. The proposed algorithm DBP Maximal Closed SPAM (MCS) showcased in the figure1, which accomplished by the direct bit position of items is manipulated. When the *min_sup* value is low, the DBP Maximal Closed SPAM evidently outperforms the previous Closed SPAM. It clears about the speedup of the algorithm, when the support value is increased.

Table 5: UCI Online Retail Data Set Sample

Invoice No	Stock Code	Description	Qty	Invoice Date	Unit Price	Cust. ID	Country
536365	85123 A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010	2.55	1785 0	UNITED KINGDOM
536365	71053	WHITE METAL LANTERN	6	01-12-2010	3.39	1785 0	UNITED KINGDOM
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010	2.75	1785 0	UNITED KINGDOM
536365	84029 G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010	3.39	1785 0	UNITED KINGDOM
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010	3.39	1785 0	UNITED KINGDOM
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010	7.65	1785 0	UNITED KINGDOM
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010	4.25	1785 0	UNITED KINGDOM
536366	22633	HAND WARMER UNION JACK	6	01-12-2010	1.85	1785 0	UNITED KINGDOM
536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010	1.85	1785 0	UNITED KINGDOM
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010	1.69	1304 7	UNITED KINGDOM
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01-12-2010	2.1	1304 7	UNITED KINGDOM
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	01-12-2010	2.1	1304 7	UNITED KINGDOM

536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01-12-2010	3.75	1304 7	UNITED KINGDOM
536367	22310	IVORY KNITTED MUG COSY	6	01-12-2010	1.65	1304 7	UNITED KINGDOM

Table 6: Characteristics of UCI Online Retail Datasets

S.No	Descriptions	Value
1	Data Set Characteristics	Multivariate, Sequential
2	No. of Instances	5,41,909
3	Number of Attributes	8

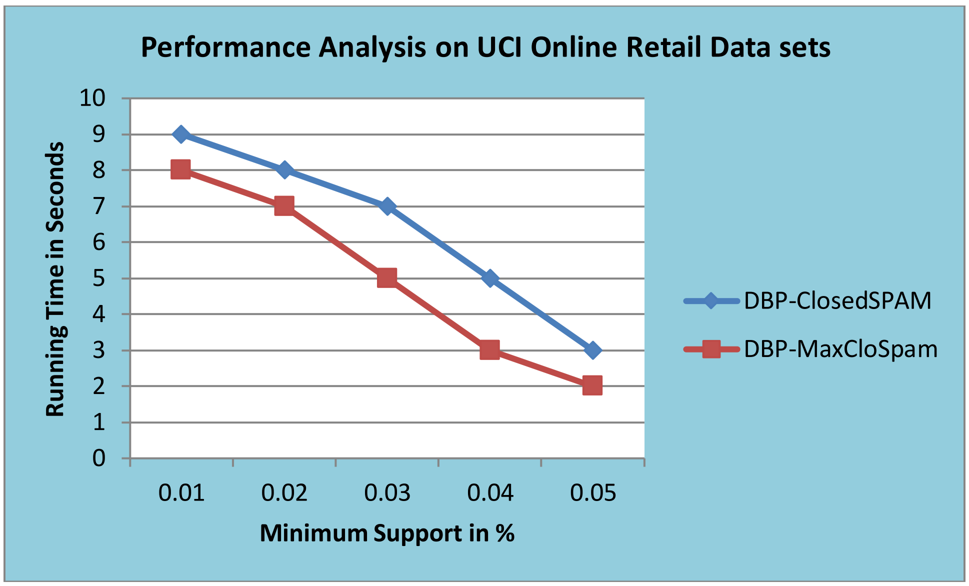

Figure1: Performance Analysis on UCI Online Retail Datasets

VIII. CONCLUSION

The proposed algorithm DBP Maximal Closed SPAM is to obtain the Maximal closed frequent sequential patterns from Sequence Database. The main challenge of sequential pattern mining depends on the size of the candidates generated and squeezes in the computations involved for the support count. This algorithm is simply extended from Direct Bit position method using SPAM.

Table 7: Output for Sample Database D (table1)

Closed Patterns	Supp.
aa	2
a c a	2
a(c d)	2
a d d	2
b c d	2
b d	2
c a d	2
(c d)d	2

The results table7 shows that the proposed algorithm is able to get the complete Maximal Closed Sequential Pattern from the given sequence database with minimum support threshold value

REFERENCES

1. *R. Agarwal and S.Arya, .Mining multiple level Association Rules to mining Multiple level Correlation to discover complex patterns. In Proc. 2012, International Journal of Computer Science, 2012.*
2. *Mabroukeh, N.R., Ezeife, C.I.: A taxonomy of sequential pattern mining algorithms, ACM Computing Surveys 43(1), 1{41 (2010)*
3. *X. Yan, J. Han, and R. Afshar.CloSpan: Mining Closed Sequential Patterns in Large Datasets. SDM'03*
4. *J. Wang and J. Han, BIDE: Efficient Mining of Frequent Closed Sequences, ICDE'04*
5. *J . Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan : Mining sequential patterns efficiently by prefix-projected pattern growth," Proc. Int'l Conf. Data Engineering (ICDE '01), pp. 215-224, Apr. 2001.*
6. *R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," Proc. Int'l Conf. Extending Database Technology (EDBT '96), pp. 3-17, Mar. 1996.*
7. *J. Pei, J. Han, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.C. Hsu, PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In ICDE'01, Heidelberg, Germany, April 2001*
8. *Nicolas Pasquier, Yves Bastide, RafikTaouil, and Lot Lakhal, "Discovering frequent closed item sets for association rules," Proceedings of the 7th International Conference on Database Theory (ICDT '99), pp. 398-416, 1999.*
9. *J. Wang, J. Han, and Chun Li, "Frequent closed sequence mining without candidate maintenance," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 8, pp. 1042-1056, Aug. 2007.*
10. *Philippe Fournier-Viger, Cheng-Wei Wu, Antonio Gomariz, Vincent S.Tseng, VMSP: Efficient Vertical Mining of Maximal Sequential Patterns, AI 2014.*
11. *K. Subramanian, E. Elakkiya, Modified Sequential Pattern Mining Using Direct Bit Position Method", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, 2016.*