An Improved Version of Closed Spam by using Direct Bit Position Method

E.Elakkiya, S.Ravichandran, S.Suresh

Abstract— ***This research paper is a new approach to Closed SPAM by Using Binary DBP Closed SPAM. The idea behind this algorithm is extending the previous DBP using SPAM. The feature DBP Closed SPAM is to retrieve large sequence database from closed Sequential Patterns The test evaluation is done on Real Datasets, experimented results shows that the algorithm DBP Closed SPAM is more efficient than the previous Closed SPAM algorithms. This is also a way to retrieve the closed frequent patterns in a sequence database by giving minimum support value.***

Index Terms - Pattern Mining, Frequent Sequential Pattern, Closed Sequential, Bit Position Method.

I. INTRODUCTION

Typically, raw-data are analyzed according to their properties and relations, can differ significantly from relational data, sequential data to graphs, cluster models, classifiers, or the combinations of these. A plethora of data mining methods and algorithms are employed to analyze different forms of raw data to ensure that the results are assured to be interpretable and understandable. The Sequential pattern mining technique is initiated by Agarwal [1]. This technique is influential to take a good decision for better solutions in the business world. Many efficient algorithms based on sequential itemset mining are readily available. But data miners envision cost-effective sequential pattern item sets from the large database. The many algorithms were created and used to retrieve information. Moreover, the difficulties are to generate candidates for mining particular frequent sequential patterns in large sequential database and the duration of execution time is the same. Instead of generating the candidates, the DBP SPAM is initiated with the feature of the base table. Finding the records that are related to closed frequent sequential patterns in the large database is ambitious as the search space extremely huge. In a sequence database, there are more than thousands of hidden patterns are stored at the same time more than Ten Thousand of possibilities for the same. This paper is especially to mine the closed frequent sequential pattern using the Direct Bit Position Method.

II. PROBLEMSTATEMENT

The taxonomies of Closed (Apriori based closed frequent itemset) [2] Algorithm works based on the Apriori algorithm along with the concept of the closed itemset lattices concept. It is traditionally based on selecting those patterns that appear in a large enough fractions of input-sequences from the database. It can be easily derived from its same supports in the super-patterns is known as supports and also it is redundant.

The Closed SPAM only generates frequently closed subsequence that does not have any supersets with the same support. Instead, it generates the mining as it is fully based on a complete set of frequent itemset mining and it is based on Apriori approach. It needs order matching is more complex. The search space of closed sequences is higher than closed sequence itemsets. It produces more than compact results without losing any information. It is sufficient to find closed itemsets.

Though most of the previous methods outfit the factors to a certain degree, the sequence ordered item is not utilized in the frequent itemset mining process. Thus, in this research, the proposed method is called DBP-Closed SPAM for mining the itemsets of closed frequent patterns of frequent sequential patterns. Moreover, pruning methods are used to retrieve and pruned by its support value and positional data item.

III.MININGPRELIMINARIES

ConsiderX = $\{x_1, x_2, x_3\ldots x_n\}$,X as a set of unique items. The X sequence of S is anorendereditem in a list, representedas S = $\{s_1, s_2, s_3\ldots s_m\}$. Them-sequence is the actual length of sequence S. In the sequences of itemsets, the brackets are ignoredwhen the element has only one element, i.e. (x_1) is written as x_1. A data item can occur many times in different sets of a sequence. A sequence X=$\{x_1, x_2, x_3..x_n\}$ is contained in another sequence Y=$\{y_1,y_2,y_3,.. y_m\}$, if there exist integers $1\leq i_1< i_2< i_3 \ldots < i_n\leq m$ such that $x_1= yi_1$, $x_2=yi_2$, …. $x_n= yi_n$. Where the sequence X is carried in another sequence ofY, the sequence X is known as subsequence of Y and Yis a super-sequence of X.it is denoted by $X\subseteq Y$. A typical sequence database D holds a numbers of set of sequences, and also have itssupport value is as the number of sequences that contain S. A frequent sequence data item is a sequence itemsets with the minimum support value (min_sup) is greater than or equal to the defined or given minimum support value.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

Dr. E.Elakkiya*, Assistant of Department of Computer Application, Alagappa University, Karaikudi, India.

Dr. S. Ravichandran, Assistant Professor, Head of the Department, of Computer Science, The H.H. Rajah's College (Autonomous), Pudukkottai, India.

Mr. S. Suresh, Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science, Coimbatore, India.

IV. PROPOSED APPROACH

An improved-version of frequent sequence pattern mining algorithm is proposed as DBP-SPAM. DBP-SPAM is to create binary digit represented item positiontable for all the sequences of the sequence database D. Let be a sequence S1=<(cd)ad>.

The item's position is founded travel over the sequence from left to right and its equivalent positions are stored. In database D, to define the binary length contain several rows in the position table value is equivalent to the length of the sequence S. The bit position table is constructed by binary digits (bit values) as 0 or 1. The presented item positions are represented as 1, when item X is existingin the i^{th} position, otherwise 0 in that position.

First, it scans the sequence database only once to retrieve the positional information and the distinct items as candidate items. It gains all frequent itemsets or length-1sequences by accomplishing their positional information. The positional data of an item i is represented by POi, consists two pairs (SD, ED), where SD is the sequence identifier, ED is the item element identifier. It is because, SD pointed out that sequence the item existencein, and EDstatesthe order of the item presented in the sequence,these depictions reserve the information of ordering the data item without any loss of sequence data. Consider a sample database in tabel1.

SD	Sequence
S1	< B (D E) B E>
S2	<B D B F>
S3	<D B E (C D E)>
S4	<C C D>
S5	< (C D E) E>

Table1: Sample Sequence Database D

The table1(A) shows theproposed algorithm is the complete frequent itemsets of Closed Pattern from the given sequence database with the minimum support threshold value.

Closed patterns	Support
B B	2
B D	3
B D B	2
B (D E)	2
B E E	2
C D E	2
C E	2
D B	3
D B E	2
D E	3
(D E) E	2

Table1(A) Output for Sample Database D

Andvalue=2 is the minimum support threshold, the positional data items as shown in table2. There is no difficulty in managing lexicographical prefix tree in this method. Hence, this is an efficient than the previous methods for closed SPAM.

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

Table2: Positional Data for a sample Sequence

To deal with the sequence S1 in the sample database D, there are three elements and four sequences as shown in the table2. If the item is present in the sequence, it is defined by 1 differently denoted by 0. It shows on the table2. Since ‘a’ is present in the sequence 1 and 3, the bit position equivalent to ‘a’ is 1010. All the sequence is constructed the same way for positional data.

To diminish the computational cost of checking bits in the position table, Itempresence table is to make with three fields, namely Item, SD and supports (sup.). Here,in addition to 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 fixed by 1, differently it is fixed by 0. Consider the item “a” in sample sequence database S since “a” is present in S1, S2, S3 the item presence table is defined as I = 1, 1, 1, 0, 0. The complete item PresentIn table is constructed like shown in thetable3.

Item	**B**	**C**	**D**	**E**	**F**
S1	1	0	1	1	0
S2	1	0	1	0	1
S3	1	1	1	1	0
S4	0	1	1	0	0
S5	0	1	1	1	0
Sup.	3	3	5	3	1

Table 3: PresentIn Table

Then the candidate items are straight from the Position-In Table and PresentIn Table. In these tables are explained without any doubt. As an alternative of achieving the candidates by inserting a data into pre-known frequent patterns, to directly achieve the proposed approach candidates using the bit position table and the PresentIn table. Let us consider that the min_sup threshold value is given by the user is 2. From, ItemPresentIn table, item “e” is pruned since the support value is 1, it less than the threshold min_sup=2 thatis sup (F)<min_sup. Now to deal with the PresentIn table is to create candidates, the first item ‘B’ and ‘C’ are considered. B = {1 1 1 0 0}, C={0 0 1 1 1}, up to this time to find (B) (C) and (“B” “C”). The operation AND is carried out in the PresentIn table values of ‘B’ and ‘C’ like, the previous DBP-SPAM.

Example1:Let’s take table1 is the input sequence database D. If the min_sup=2, the closed sequential itemsets are CS={(BB):2, (BD):3,(BDB):2, B(DE):2,(BEE):2,(CDE):2,(CE):2, (DB):3, (DBE):2, (DE):3, (DE)E: 2} from the total FS 13 sets of sequences={(BB):2,(BC):3,(BCB):2,B(DE):2,(BE):2, (BEE):2, (CD):2, (CDE):2, (CE):2, (DB):3, (DBE):2, (DE):3, (DE)E :2}

Example2:Consider another sample databaseshown in table4, referred as D where the perspective of unambiguous. The alphabetic order is engaged as the lexicographical order. Whenmin_sup=2, the Closed SequentialFrequent Pattern itemsets are (CS) = {(BG) E: 2, (FB):3, (FBC):2} while the corresponding Frequent Sequence (FS) set has 16 sequences. CS has the exact data itemsets as FS, but embraces much less frequent patterns.

SID	Sequence
1	<(BG)EFB>
2	<FBC>
3	<F(BCG)(CEF)>

Table 4: Sample Database

V. ALGORITHM OF DBP CLOSEDSPAM

In step1, the database scanned to build the PresentIn Table for 1-sequence. Step2, begin with 2-sequence AND 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 (CS) Closed Frequent Sequences are retrieved from the complete (FS) Frequent Sequence by comparing the each FS itemset.

VI. PSEUDO CODE FORDBPCLOSEDSPAM

DBPClosedSPAM (SDB, min-sup)

```
INPUT:SDB – Sequence Database, min-sup– Minimum
Support value,
OUTPUT: Closed Sequential pattern itemsets
BEGIN:
//BitPositions sequences
// PresentIn sequences and set Scount as 0
For each [SD, S]<= D begin
For each Element Sj of S begin
For each item i<=Sj begin
If PresentInItem (i) = 0, Mark PresentInItem (i =1)
Set jthbit in POi(i) =1
End for
End for
End for
Patterns= IS_PatternItems(PresentIn, min-sup)
closedPatterns=closedDBP_SPAM(Patterns,min-sup)
END
Function ClosedDBP_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 closed or not
//check pattern (i) is super or Sub sequence of Pattern(j)
If Patterns (i) is like Patterns(j) then
//compare support values of both patterns
 If supp(i)=supp(j) then
Next
Else
Return Patterns(i)
End if
Else
Return Patterns(i)
Endif
End For Loop
End For Loop
END
Function IS_PatternItems(PresentIn,min-sup)
INPUT: PresentIn table, min-sup
BEGIN:
For Eachitemsi <= PresentInTable
Form base itemsets
// by applying AND operation
If base Items>= min-sup
//Store base itemsets in base table
 End for
For each Items K <= BaseTable
Fetch POi 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 DBP-SPAM algorithm is executed on a system comprising of 2.66 GHz Pentium dual core processor machine with a 2 GB memory running on Microsoft 7 ultimate operating system.The algorithm is written in visual C# programming.The experimental evaluation is performed by KDDCup2000 is shown on table5 and the characteristics intable6.

ClickStreams (Items)	ItemID
'buffer_overflow.'	117
'ftp_write.'	118
'guess_passwd.'	119
'imap.'	120
'ipsweep.'	121
'land.'	122
'loadmodule.'	123
'multihop.'	124
'neptune.'	125
'nmap.'	126
'normal.'	127
'perl.'	128

Table 5: Sample Real Dataset of Gazelle Web View

Algorithm	Data Size in K				
	4K	5K	6K	7K	8K
	Running Time in Seconds				
DBP-SPAM	30	41	55	72	122
Parallel DBP-SPAM	15	21	26	34	57
DBP-ClosedSPAM	28	36	51	68	120

Table6: KDD Cup 2000 (Gazelle) Dataset Characteristics

The figure1 shows the performance ofthe proposed algorithm executed and discovered the closed sequential patterns effectively eventhe running time is compared to Gazelle real- world dataset. The minimum support value is replaced from 1% to 5%.

The experiments are accomplished changeable min-sup values. The performance of proposed algorithmDBP Closed SPAM to signify, the foremost reason for this high-speed execution depend on the pruning technique, which achieved by the direct bit position of items ismanipulated.When the min-sup value is low, the DBP Closed SPAM is evidently outperforms the previous Closed SPAM compares with Clospan and Clasp. From the chart, it clears about the speedup of the algorithm, when the support is increased. The table 7 and 8 shows the comparison of Gazelle Datasets.

Sno	Particulars	Value
1	Number of Sequences	77,512
2	Distinct Items	3,340
3	Sequences' Average Length	4.62
4	Items Standard Deviation	6.07

Table 7: Comparison Gazelle Datasets with Min_Sup

Algorithm	Min_sup				
	0.01	0.02	0.03	0.04	0.05
	Running Time in Seconds				
DBP-SPAM	72	52	23	18	14
Parallel DBP-SPAM	31	16	9	6	4
DBP-ClosedSPAM	30	15	8	5	4

Table8: Comparison Gazelle Datasets with Data Size

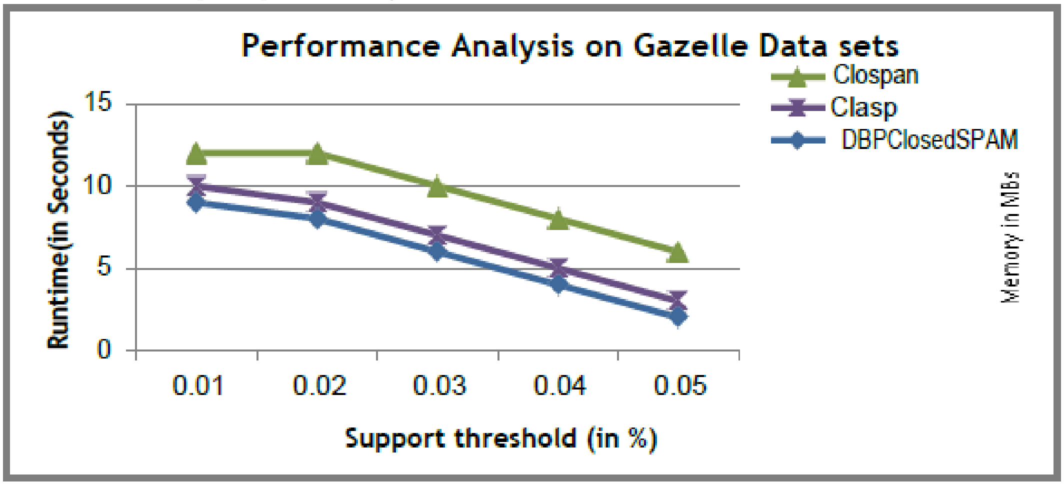

Figure 1: Performance Analysis on Gazelle Data sets

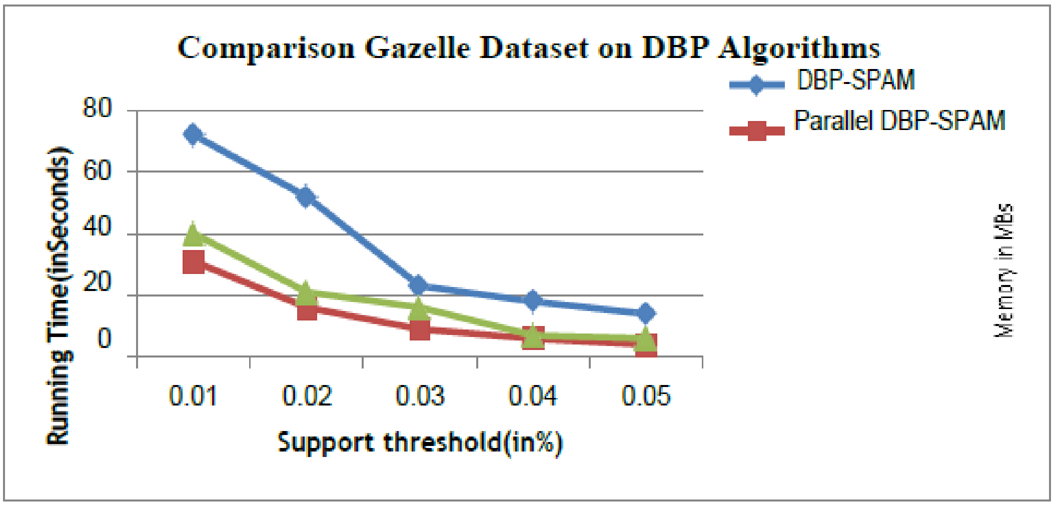

Figure 2: Comparison Gazelle Dataset on DBP Algorithms

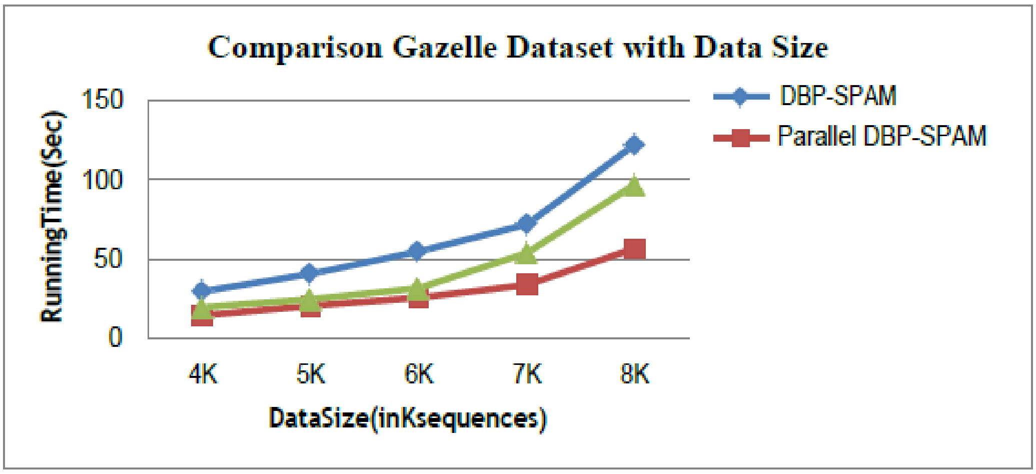

Figure 3: Comparison Gazelle Dataset with Data Size

The resultant table data is shown in figure2 and 3. The analysis is made on these Direct Bit Position methods. From the figure2 and 3 it clearly shows that when the data size is huge, Parallel DBP-SPAM is the best of mining frequent patterns. The closed pattern is the ideal and maximum lengths of patterns are few in most databases. Obviously, the closed pattern is almost very much similar to Parallel DBP-SPAM.

VIII. CONCLUSION

The performance of proposed algorithm DBP Closed SPAM is to obtain the closed frequent sequential itemsets of patterns from Sequential Database. The primary challenge and objective of the sequential pattern mining is depends on the size of the candidate itemsetsaregenerated and it manage to get the computations is involved in the support value count.This algorithm is simply extended with an efficient method from the Direct Bit Position method using SPAM. According to the experimental evaluation, the comparative study has been made with original Gazelle datasets, and it is clearly shown that, the parallel algorithm is the best one for frequent pattern mining compared with execution time. The huge datasets like gazelle have always needed to process data with the short amount of execution time. The DBP Closed SPAM with Parallel algorithm with more processing components for better performance as future work.

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. V. PurushothamaRaju and G.P. SaradhiVarma, Mining Closed Sequential Patterns in Large SequenceDatabases,IJDMS, Vol7, No.1, February 2015
3. 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.
4. C. C. Aggarwal, J. Han (Eds.), Frequent Pattern Mining, 65 DOI 10.1007/ 978-3- 319-07821-2_3, Springer International Publishing Switzerland2014.
5. Antonio Gomariz, Manuel Campos, Roque Marin, and Bart Goethals, “ClaSP: An efficient algorithm for mining frequent closed sequences,” PAKDD 2013, LNAI 7818, Part I, pp. 50–61, 2013.
6. S.MuthuSelvan, KS. Sundaram, “A survey of Sequence Patterns in Data Mining Techniques. International Journal of Applied Engineering Research2015.
7. Fournier-Viger, P., Faghihi, U., Nkambou, R., MephuNguifo, and E.: CMRules: Mining Sequential Rules Common to Several Sequences. Knowledge-based Systems, Elsevier, Vol. 25, No. 1, pp. 63-76(2012).
8. Pasquier, Yves Bastide, RafikTaouil, and Lot Lakhal, “Discovering frequent closed itemsets for association rules,” Proceedings of the 7th International Conference on Database Theory (ICDT '99), pp. 398-416, 1999.
9. E. Elakkiya,S. Ravichandran, “Max-closed by using Direct Bit Position Method”, Global Journal of Engineering Science and Researches(GJESR), ISSN (Online): 2348-8034, 2019.
10. K. Subramanian, E. Elakkiya, “ A New Parallel Algorithmic approach for sequential pattern mining using Binary Representation”, International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS), ISSN (Online): 2321-7782,2016.

AUTHORS PROFILE

Dr. E.ELAKKIYA received the Ph.D degree in Computer Science from J.J. College of Arts and Science, Pudukkottai, India. She is currently working as a Teaching Assistant of Department of Computer Application, Alagappa University, Karaikudi, India. Her research interests include Data Mining, Big Data

Dr. S. RAVICHANDRAN received the Ph.D degree in Bharathidasan University, Trichirappalli, India. He is currently working as aAssistant Professor, Head of the Department, Department of Computer Science, The H.H. Rajah’s College (Autonomous), Pudukkottai, India. His research interests include Data Mining, Big Data Analytics, and Image Processing. He had published 13 International Journals.

Mr. S. SURESH received theM.Sc Degree in BharathidasanUniverisity&M.Phil degree in Vinayaga Missions University, Salem, India. He is currently working asAssistant Professor, Department of Computer Science, Sri Krishna Arts and Science, Coimbatore India. He also got through the SETExam held on Oct 2012 byBharathiar University, Coimbatore and his research interests are Data Mining, Big Data Analytics, and Image Processing, IOT.