MINING CLOSED REGULAR PATTERNS IN DATA STREAMS

M.Sreedevi[1] and Dr.L.S.S.Reddy[2]

[1]Department of Computer Science and Engineering, K L University, Guntur,
Andhra Pradesh, India
msreedevi_27@kluniversity.in
[2]Department of Computer Science and Engineering, LBR College of Engineering,
Mylavaram, Andhra Pradesh, India
director@lbrce.ac.in

ABSTRACT

Mining regular patterns in data streams is an emerging research area and also a challenging problem in present days because in Data streams new data comes continuously with varying rates. Closed item set mining gained lot of implication in data mining research from conventional mining methods. So in this paper we propose a narrative approach called CRPDS (Closed Regular Patterns in Data Streams) with vertical data format using sliding window model. To our knowledge no method has been proposed to mine closed regular patterns in data streams. As the stream flows our CRPDS-method mines closed regular itemsets based on regularity threshold and user given support count. The experimental results show that the proposed method is efficient and scalable in terms of memory and time.

KEYWORDS

Data Streams, Regular patterns, closed regular patterns, transaction sliding window.

1. INTRODUCTION

Mining data streams efficiently is a challenging area in a mining research because new data arrives and old data is overdue with rapid speed. Data streams emerging class of applications in recent years which is often continuous, unbounded, high speed and data distribution as time changes [1]. Data streams can be classified into two types, they are a) off line data streams b) on line data streams. Web log reports, queries on data warehouse are examples for offline data streams. Network transactions, bank transactions, sensor data, etc are examples for online data streams. Mining data stream requires fast, real time processing in order to keep up with high speed data arrival and results must be attracted with in short response time. Similarly multiple scans in data streams are not adequate. Recently, Tanbeer et al. [8] introduced a new problem of discovering regular patterns that follow temporal regularity or the occurrence behaviour of a pattern. A pattern which is derived from database based on user given regularity threshold is called a regular pattern. Regularity of the item plays an important role in mining process. For example in retail market some products have demand, and it is essential to know how regularly the products are sold rather than number of items sold. They also introduced the same problem in data streams.

At present closed itemset mining gained a lot of significance than traditional frequent data mining methods. Literature survey shows that there are many methods derived to mine closed itemsets in

DOI : 10.5121/ijcsit.2013.5114

data mining research. Gupta et al., proposed a method CLICI to mine closed itemsets on data streams using formal concept analysis in landmark window model [9]. Pramod. S and Vyas. O. P proposed an algorithm to mine frequent item sets for on line data streams using prefix tree based structure. In this method the items are to be arranged in sorting order in every transaction. In this process it takes additional time to arrange every transaction items in sorted order, similarly it consumes memory also. So in this paper we propose a new method called CRPDS (Closed Regular Patterns in Data Streams) to mine closed regular patterns on data streams using vertical data format in sliding window model. The main idea of our proposed method is to develop a simple, powerful, that captures data from data streams into window using sliding window mechanism to find closed regular itemsets. Sliding window model contains the fixed number of transactions in the window, which lead to maintain constant transaction processing time. However, this model cascade the undersized monitoring of continuous changes of data streams.

The rest of the paper is organized as follows. Section 2 describes the related work; section 3 describes problem definition of closed regular pattern mining. The method CRPDS to mine closed regular patterns on data streams using vertical data format with sliding window model is discussed in section 4. In Section 5 we describe experimental results and finally conclude the paper in section 6.

2. RELATED WORK

Discovering interesting patterns in data streams is taxing problem in recent existence. Guohui Li et al., proposed a method for mining frequent patterns in an arbitrary sliding window of data streams by using time decaying model to discriminate patterns of recent transactions with old transactions [2]. In this process as the stream flows SWP tree captures the contents of stream data and infrequent patterns are deleted. OPFI – Stream algorithm using prefix tree data structure to mine frequent itemsets with sliding window technique over data streams proposed byKun li et al., in [3]. T.Calders et al., in [11] proposed Optimized incremental algorithm for mining frequent itemsets in data streams. Leung et al., proposed DStree structure for mining the frequent sets in data stream using sliding window [4]. In this process the transactions are arranged based on canonical order specified by user prior to construction of tree later mining process start on this tree data. Giannella et al., derived FP-stream approach to mine frequent patterns in data streams at multiple time granularities [5]. CPS-tree(Compact pattern stream tree) captures recent stream content data and mine complete set of frequent patterns from high speed data stream over sliding window model by avoiding obsolete and old transactional data[12][13].

The occurrence frequency may not be sufficient and temporal regularity in occurrence behaviour of item also required for many data stream application like stock market analysis, network monitoring analysis etc. Traditional frequent mining methods fail to cover occurrence behaviour of itemsets because these methods focused mainly on frequency of item occurrence. Literature survey shows that mining regular patterns in statistical data bases have been addressed. RPS-tree for discovering regular patterns over data streams proposed by Tanbeer S.K. et al., in [6] and in this process sliding window mechanism and tree based structure is used to capture regular itemsets in data streams. VFDT is one algorithm to mine decision trees from continuously changing data in data streams and similarly CVFDT is another algorithm by reapplying VFDT algorithm on moving window every time proposed by Geoff Hulten et.al., in [14]. Vijay Kumar et al., VDSRP method to generate complete set of regular patterns over a data stream at a user given regularity threshold using vertical data format [7].

3. PROBLEM DEFINITION

Let I = $\{i_1, i_2, i_3, \ldots, i_n\}$ be a set of items. A set X = $\{i_1, i_2, \ldots, i_q\} \subseteq I$, where $l \leq q$ and $l, q \in [1, n]$ is called an item set or a pattern. A transaction t = (tid, Y) is a tuple where tid is a transaction

Id and Y is a pattern. The set of transactions T = {t_1, t_2, …, t_m} is a transactional database DB over I.

3.1. Definition 1 (Data Stream)

A Data Stream DS can be defined as a continuous flow of transactions with high speed. Assume DS = {t_1, t_2, $t_{3,}$ t_4, … , $t_{m, ...}$}, t_i is the i^{th} arrived transaction and i∈[1, m]. The window W contains a set of transactions which are arrived from i^{th} arrival to j^{th} arrival of transactions, where i is always less than j. The size of window is |W| = j – i+1, i.e, the number of transactions arrived between i^{th} and j^{th} arrival of transactions.

3.2. Definition 2 (Transaction Sliding window T_{SW} of DS)

Transaction sliding window T_{SW} contains a fixed number of transactions in data streams. Slide of window introduce and expires the slide–size i.e, 1 ≤ slide-size ≥ |W|, the transactions into and from the transaction window. If X occurs in t_j, the transaction-id of X is represented as t_j^X, j ∈ [1 , |W|]. Therefore T_w^X = {t_j^X , . , . , . , t_k^X} j, k ∈[1,|W|] and j ≤ k for the set of all transaction-ids where X occurs in transaction window T_{SW}.

3.3. Definition 3 (period of X in T_{SW})

Let t_j^X and t_{j+1}^X are two consecutive transaction-ids in transaction sliding window T_{SW}. The number of transactions between t_j^X and t_{j+1}^X is defined as a period of X, say p^X where $p^X = t_{j+1}^X - t_j^X$, j∈(1, |W|). We consider the first transaction is t_f which is null transaction i.e t_f = 0 and last transaction is t_l which is last transaction in the T_{SW}.

For example, consider the Table1 of stream data where first transaction sliding window T_{SW1} consists of eight transactions i.e., from tid-1 to tid-8. Set of transactions of window T_{SW1} where pattern b appears in the transactions (2, 3, 5) and the periods for pattern b are {(2 – t_f) = 2, (3-2) = 1, (5-3) = 2, (t_l – 5) = 3}. Similarly pattern (b c) appears in transactions (1, 5, 7) and the periods for (b c) are {(1-t_f) = 1, (5-1) = 4, (7-5) = 2, (t_l-7) = 1} where t_f = 0 and t_l = 8. The items b and (b c) are provided the information about their occurrence periods of the window T_{SW}.

3.4. Definition 4 (Regularity of X in T_{SW})

Let p_w be the set of all periods of X in the T_{SW} i.e $p_w(x)$ = {p_1^X,……p_q^X} where q is the highest transaction number for X appears in the window. The regularity of X in the window is defined as reg (X) = max({p_1^X,……p_q^X}). For instance regularity of pattern b in T_{SW1} is 3 i.e., P_W(b) = max(2, 1, 2, 3) = 3 and regularity of pattern (b c) is 4 i.e., P_w(b c) = max (1, 4, 2, 1) = 4. We say the pattern is regular if it occurs in the specified period other wise it is not regular pattern. So the regularity of pattern depends on the occurrence behaviour of pattern in the T_{SW}.

3.5. Definition 5 (Closed itemset)

The itemset X is closed in transaction sliding window T_{SW} if there exist no proper superset S has same support as X in T_{SW}. For example α = (a_1, a_2, …, a_m) is one itemset and β = (b_1, b_2, …, b_n) is another itemset. Let α is subset of β (i.e., α ⊆ β), therefore β contains α. The support count of α is denoted as sup(α) and support count of β is denoted as sup(β) and sup(α) > sup(β), itemset α is closed itemset for user given minimum support threshold.

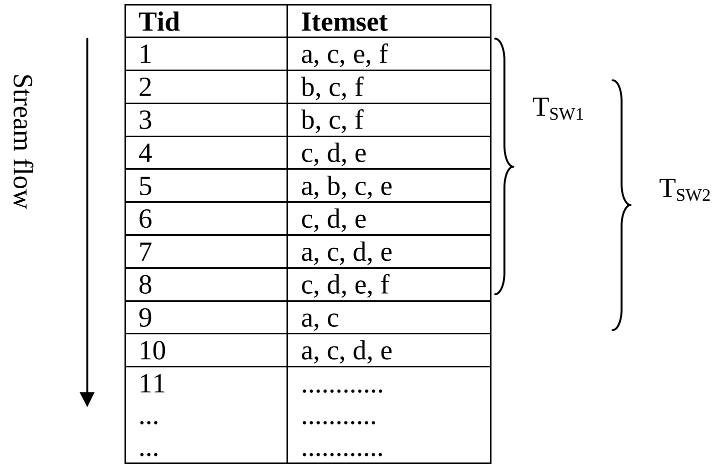
3.6. Definition 6 (Closed regular itemset)

Let X = {x_1, x_2, …., x_m} be one set of regular items and Y = {y_1, y_2, ….., y_n} be a another set of regular items where $X \subseteq Y$, X is subset of Y and Y is superset of X. support count of Y must not be greater than support count of X for user given minimum support threshold S i.e., sup(X) > sup(Y), X is closed regular itemset.

4. MINING CLOSED REGULAR PATTERNS IN DATA STREAMS

Traditional method like Apriori and FP-growth algorithms used horizontal data format to mine interesting patterns. Instead of using traditional methods our proposed method uses vertical data format with sliding window model, where vertical data format requires one database scan which contains fixed number of transactions in a sliding window every time. We consider the data base which is in [7] is our running example to mine closed regular patterns. CRPDS method is implemented in two phases i.e., regular patterns are mined in first phase and closed regular patterns are mined in the second phase.

Table 1. Data Stream



Tid	Itemset
1	a, c, e, f
2	b, c, f
3	b, c, f
4	c, d, e
5	a, b, c, e
6	c, d, e
7	a, c, d, e
8	c, d, e, f
9	a, c
10	a, c, d, e
11	....

Let Table1 contains series of transactions of data stream DS. Data stream contains transaction-id and set of items corresponding to transaction-id i.e., tid and itemset. Consider the transaction sliding window T_{SW} of size 8 i.e., |W| = 8 i.e., from tid-1 to tid-8. Convert the T_{SW1} transactions in to vertical format (i.e., itemset, tid number) and find the periodicity of each item P^X and consider maximum periodicity value as a regularity of an item. Assume maximum regularity threshold λ = 4 and items which are having regularity is less than or equal to λ are regular items which are shown in Table2.

Phase 1

Input: T_{SW} in DS, λ
Output: Set of regular items
Procedure

1. Consider T_{SW} of DS, |W| = 8
2. Convert T_{SW} into Vertical data format.
3. Let $X_i \in T_{SW}$, $X_i \subseteq$ k-itemset.
4. $P^X{}_i = 0$ for all X_i
5. For every X_i calculate periodicity

```
6.  $P^X_i = P^X_{i+1} - P^X_i$
7.  $reg(X_i) = max(P^X_i)$
8.  if $reg(X_i) < = \lambda$
9.          $X_i$ is regular item set
10. Else
11.     Delete $X_i$
12. Repeat step3 to step11 for i+p itemsets (p = 1, 2, 3, ...)
```

Table 2. T_{SW1} in vertical data format

Itemset	tid
a	1, 5, 7
b	2, 3, 5
c	1, 2, 3, 4, 5, 6, 7, 8
d	4, 6, 7, 8
e	1,4, 5, 6, 7, 8
f	1, 2, 3, 8

In Phase I regular itemsets are mined in T_{SW1} of Data stream DS. First we convert horizontal transactions of T_{SW1} into vertical data format of size |W| = 8. X_i is an itemset and P^X_i is periodicity of itemset X_i. We calculate periodicity P^X for each itemset and take maximum periodicity as the regularity of an itemset.

Table 3. T_{SW1} with itemset, P^X and Reg

Itemset	tid	P^X	Reg
a	1, 5, 7	1, 4, 2, 1	4
b	2, 3, 5	2, 1, 2, 3	3
c	1, 2, 3, 4, 5, 6, 7, 8	1, 1, 1, 1, 1, 1, 1, 1, 1	1
d	4, 6, 7, 8	4, 2, 1, 1	4
e	1,4, 5, 6, 7, 8	1, 3, 1, 1, 1, 1	3
~~f~~	1, 2, 3, 8	1, 1, 1, 5	~~5~~

Itemsets of T_{SW1} with their periodicities and their regularities are present in Table 3. For example itemset (d) is appeared in transactions (4, 6, 7, 8) and their periodicity values $P^X(d) = \{ 4, 2, 1, 1\}$.Regularity value is 4 i.e., max(4, 2, 1, 1) = 4. In our running example the minimum regularity threshold is $\lambda = 4$. The itemsets which are having the regularity is less than or equal to minimum regular threshold are regular items. Therefore items {a, b, c, d, e} are regular itemsets and itemset {f} is not a regular itemset which are shown in Table 3.

In Table 4 the itemsets {(a c), (a e), (b c),(c d), (c e), (d e)} having the regularity which is less than or equal to λ are regular itemsets and rest of them are not regular itemsets. Similarly 3-itemsets, 4-itemsets, and so on can be mined from the previous regular itemsets.

In the second phase closed regular itemsets are mined from previously mined regular itemsets which are shown in table 5 and table 6.

Table 4. Two itemsets with P^X and Reg

Itemset	tid	P^X	Reg
(a b)	5	5, 3	~~5~~
(a c)	1, 5, 7	1, 4, 2,1	4
(a d)	7	7, 1	~~7~~
(a e)	1, 5, 7	1, 4, 2, 1	4
(b c)	2, 3, 5	2, 1, 2, 3	3
(b e)	5	5, 3	~~5~~
(c d)	4, 6, 7, 8	4, 2, 1, 1	~~4~~
(c e)	1, 4, 5, 6, 7, 8	1, 3, 1, 1, 1, 1	3
(d e)	4, 6, 7, 8	4, 2, 1, 1	~~4~~

Phase II

Input : Regular itemsets, S

Output: Complete set of closed regular itemsets.

```
1. Let $X_i \subseteq I$ is a regular p-item set
2. Let $X_j \subseteq I$ is a regular p+k item set
   Where k varies from 1 to n
3. $X_i \subseteq X_j$ for all i <= j
4. Find support-counts of $X_{i,}$ and $X_j$ i.e., Sup($X_i$) and Sup($X_j$)
5. For User given minimum support S
6. If Sup($X_i$) > Sup($X_j$)
7.     $X_i$ is closed-regular item set
8. Else
9.     Delete $X_i$
```

Table 5. Regular one itemsets with Support

Itemset	tid	Reg	Sup
a	1, 5, 7	4	~~3~~
b	2, 3, 5	3	~~3~~
c	1, 2, 3, 4, 5, 6, 7, 8	1	8
d	4, 6, 7, 8	4	4
e	1,4, 5, 6, 7, 8	3	6

One itemsets with their regularity and support count values are shown in table 5.Assume S = 4, itemsets (c), (d), (e) are satisfied the specified minimum support value S and itemsets (a), (b) are not satisfied the minimum support value.

Two itemsets { (a c),(a e), (b c)} have not been satisfied S and itemsets {(c d),(c e),(d e)} are satisfied S. Apply closed property on the itemsets which have been satisfied regularity and support , Support count of itemset *c* is greater than support count of itemsets {(c d),(c e)}. So itemset *c* is a closed regular itemset.

Table 6. Regular two itemsets with support.

Itemset	tid	Reg	Sup
(a c)	1, 5, 7	4	~~3~~
(a e)	1, 5, 7	4	~~3~~
(b c)	2, 3, 5	3	~~3~~
(c d)	4, 6, 7, 8	~~4~~	~~4~~
(c e)	1, 4, 5, 6, 7, 8	3	6
(d e)	4, 6, 7, 8	~~4~~	~~4~~

Table 7. Regular three itemsets with reg and sup.

Itemset	tid	Reg	Sup
(a b c)	5	5	1
(a b d)	----	----	~~---~~
(a b e)	5	5	1
(a c d)	7	~~7~~	1
(a c e)	1, 5, 7	4	3
(a d e)	7	7	1
(b c d)	----	----	---
(b c e)	5	5	1
(c d e)	4, 6, 7, 8	4	4

The itemset (c d e) satisfied the regularity and support, the support count of itemset (c e) is greater than the support count of itemset (c d e), so itemset (c e) is closed regular itemset and itemsets {(c d), (d e)} are not closed regular itemsets. Like this three itemsets, four itemsets and so on can also be mined until no closed regular itemsets found.

5. EXPERIMENT RESULTS

In this section we produced our results for closed regular patterns in data streams. We used java to develop our algorithm with the system configuration of 2.66 GHz CPU with 2GB main memory on windows XP Operating system. We applied our mining process on Kosarak (real data set) and T10I4D100K (synthetic data set).These data sets are frequently used in frequent pattern mining experiments which are developed at IBM Almaden quest research group and which are obtained from http://cvs.buu.ac.th/mining/datasets/synthesis_data and UCI machine repository (University of California,Irvine, CA). The real data set provided by Ferenc Bodan which contains click stream data of Hungarian on-line news portal.

We used T10I4D100K and Kosarak datasets with different regularity and support values to compare our results with RP-tree that finds only regular itemsets. T10I4D100K dataset contains 870 items with average length of10.10 of 1,00,759 total transactions. Kosarak dataset contains 41,270 items with average transaction length of 8.10 of 9,90,000 transactions. To produce our results we consider 100K and 500K size of T10I4D100K and kosarak datasets which are shown in figure1 and figure2 respectively. Experimental results shown that higher max-reg() and min-sup() values longer the time required which are exposed in both the graphs.

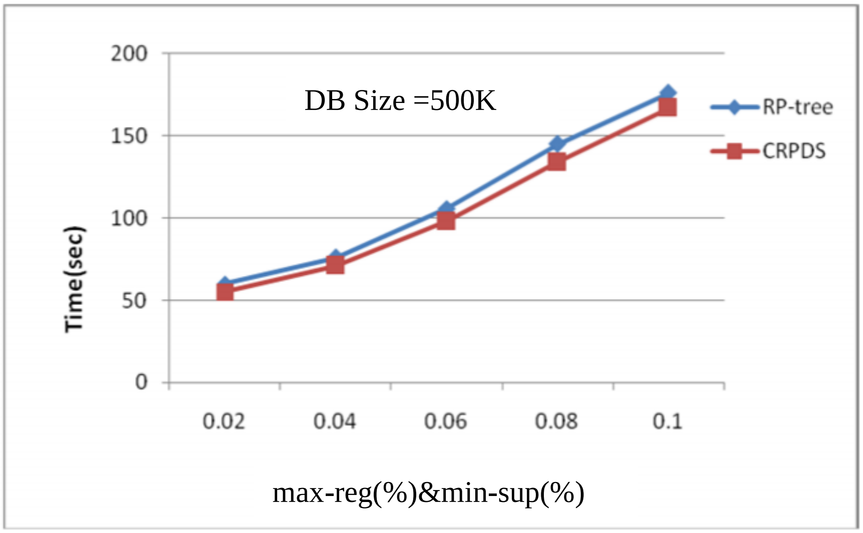


Figure 1. Execution time over Kosarak

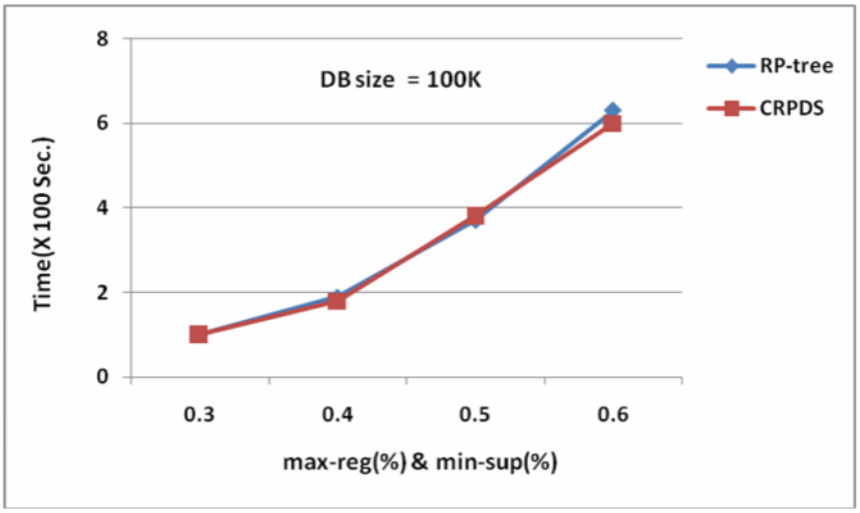


Figure 2. Execution time over T10I4D100K

6. Conclusion

Closed regular pattern mining in data streams is completely a new approach in data mining applications. We Proposed CRPDS algorithm to mine closed regular patterns using vertical data format with sliding window model. The advantage of our proposed algorithm is it requires simple operations like addition, subtraction, arrays etc. Our experimental results have shown the effectiveness of CRPDS in terms of execution time.

Acknowledgements

We are very much thankful to Sri G.Vijay Kumar, Associate professor in Department of Computer science and Engineering , K L University, who Supported and contributed towards fulfilment of our work.

REFERENCES

[1] Jiang, N., Gruenwald, L., (2006) " Research issues in data stream association rule mining." SIGMOD Record 35(1) ,pp14-19.

[2] Guohui Li, Hui chen, Bing Yang and Gang chen , (2008)" Mining frequent patterns in an arbitrary sliding wnidow over data streams", DASFAA,LNCS 4947, pp 496-503.

[3] Kun Li, Yongyan wang, Manzoor Elahi, Xin-Li and Hongan wang (2008-september) "Mining recent frequent itemsets in data streams with optimistic pruning" Springer, ECME PKDD.

[4] Leung, C.K.S, Khan, Q.I.,(2006 December) "DStree: A tree structure for mining of frequent sets from data streams. ICDM , IEEE press, Los Alamitos, pp 928-932.

[5] Giannell.C, Han,J., Pei,j., Yan.X.,Yu.Ps., (2004) "Mining frequent patterns in data streams at multiple time granularities." In Data Mining : Next generation challenges and future directions, AAAI/MIT Press, pp 191-212.

[6] S.K. Tanbeer, C. F. Ahmed, B.S. Jeong, and Y.K. Lee, (2010)"Mining Regular Patterns in data streams", DASFAA, Springer, Part I, LNCS 5981, pp 399-413.

[7] Vijay Kumar, G., Sreedevi, M., Pavan kumar, N.V.S.,(2012) " Mining Regular Patterns in data streams using Vertical format", IJCSS volume 6 Issue 2, pp142-149.

[8] S.K. Tanbeer, C. F. Ahmed, B.S. Jeong, and Y.K. Lee, (2008) "Mining Regular Patterns in Transactional Databases", IEICE Trans. On Information Systems, E91-D, 11, pp. 2568-2577.

[9] Anamika Gupta, Vasudha Bhatnagar, and Naveen Kumar (2010) "Mining closed itemsets on data streams using formal concept analysis", pp 285-296.

[10] S.pramod and O.P.vyas (2010) "Frequent Itemset mining over transactional data streams using Item-Order-Tree" IJCSE vol 2 no 8 pp2598-2601.

[11] T.Calders,N.Dexters, JJM. Gillis and B.Geothals (2012) "Mining frequent itemsets in a stream" Information Systems , Elsevier.

[12] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, Byeong-Soo Jeong, Young-Koo Lee (2009), "Sliding window- based frequent pattern mining over data streams" Information sciences 179, pp 3843-3865.

[13] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, Byeong-Soo Jeong, Young-Koo Lee (2008) " Efficient frequent pattern mining over data streams" ACM 978-1-59593-991.pp 1447-1448.

[14] Geoff Hulten, Laurie spencer, and Domingos (2001)" Mining time changing data streams" ACM-01.USA, pp 97-106.