

Yun Chi · Haixun Wang · Philip S. Yu ·
Richard R. Muntz

Catch the moment: maintaining closed frequent itemsets over a data stream sliding window

Received: 6 August 2004 / Revised: 13 September 2005 / Accepted: 24 September 2005 /
Published online: 15 March 2006
© Springer-Verlag London Ltd. 2006

Abstract This paper considers the problem of mining closed frequent itemsets over a data stream sliding window using limited memory space. We design a synopsis data structure to monitor transactions in the sliding window so that we can output the current closed frequent itemsets at any time. Due to time and memory constraints, the synopsis data structure cannot monitor all possible itemsets. However, monitoring only frequent itemsets will make it impossible to detect new itemsets when they become frequent. In this paper, we introduce a compact data structure, the *closed enumeration tree* (CET), to maintain a dynamically selected set of itemsets over a sliding window. The selected itemsets contain a boundary between closed frequent itemsets and the rest of the itemsets. Concept drifts in a data stream are reflected by boundary movements in the CET. In other words, a status change of any itemset (e.g., from non-frequent to frequent) must occur through the boundary. Because the boundary is relatively stable, the cost of mining closed frequent itemsets over a sliding window is dramatically reduced to that of mining transactions that can possibly cause boundary movements in the CET. Our experiments show that our algorithm performs much better than representative algorithms for the state-of-the-art approaches.

Keywords Data streams · Sliding window · Closed frequent itemset · Incremental learning

Y. Chi (✉) · R. R. Muntz
Department of Computer Science, University of California, Los Angeles, CA 90095, USA
E-mail: {ychi, muntz}@cs.ucla.edu

H. Wang · P. S. Yu
IBM Thomas J. Watson Research Center, Hawthorne, NY 10532, USA
E-mail: {haixun, psyu}@us.ibm.com

1 Introduction

1.1 Motivation

Data streams arise with the introduction of new application areas, including ubiquitous computing and electronic commerce. Mining data streams for knowledge discovery is important to many applications, such as fraud detection, intrusion detection, trend learning, etc. In this paper, we consider the problem of mining closed frequent itemsets on data streams. Mining frequent itemset on static data sets has been studied extensively. However, in many applications, mining frequent itemsets on data streams is needed. For example, for a commercial web site, web pages that are usually visited together by different users reveal important information on user behavior and user interests. In many cases, this information is contained in large-volume, unbounded user click streams. As another example, frequent itemsets can help traffic measurement and intrusion detection in high-speed, real-time network traffic data. Data streams have posed new challenges. First, data streams are continuous, high-speed, and unbounded. Archiving everything from streams is impossible, not to mention mining association rules from them using algorithms that require multiple scans. Second, the data distribution in streams are usually changing with time, and very often people are interested in the most recent patterns.

It is thus of great interest to mine itemsets that are *currently* frequent. One approach is to always focus on frequent itemsets in the most recent window. A similar effect can be achieved by exponentially discounting old itemsets. For the window-based approach, we can immediately come up with two naive methods:

1. Regenerate frequent itemsets from the entire window whenever a new transaction comes into or an old transaction leaves the window.
2. Store every itemset, frequent or not, in a traditional data structure such as the prefix tree, and update its support whenever a new transaction comes into or an old transaction leaves the window.

Clearly, method 1 is not efficient. In fact, as long as the window size is reasonable, and the concept drifts in the stream is not too dramatic, most itemsets do not change their status (from frequent to non-frequent or from non-frequent to frequent) often. Thus, instead of regenerating all frequent itemsets every time from the entire window, we shall adopt an *incremental* approach.

Method 2 is incremental. However, its space requirement makes it infeasible in practice. The prefix tree [1] is often used for mining association rules on static data sets. In a prefix tree, each node n_I represents an itemset I and each child node of n_I represents an itemset obtained by adding a new item to I . The total number of possible nodes is exponential. Due to memory constraints, we cannot keep a prefix tree in memory, and disk-based structures will make real time update costly.

In view of these challenges, we focus on a *dynamically selected* set of itemsets that are (i) informative enough to answer at any time queries such as “what are the (closed) frequent itemsets in the current window”, and at the same time, (ii) small enough so that they can be easily maintained in memory and updated in real time.

The problem is, of course, what itemsets shall we select for this purpose? To reduce memory usage, we are tempted to select, for example, nothing but frequent

(or even closed frequent) itemsets. However, if the frequency of a non-frequent itemset is not monitored, we will never know when it becomes frequent. A naive approach is to monitor all itemsets whose support is above a reduced threshold $\text{minsup} - \epsilon$, so that we will not miss itemsets whose current support is within ϵ of minsup when they become frequent. This approach is apparently not general enough.

In this paper, we design a synopsis data structure to keep track of the boundary between closed frequent itemsets and the rest of the itemsets. Concept drifts in a data stream are reflected by boundary movements in the data structure. In other words, a status change of any itemset (e.g., from non-frequent to frequent) must occur through the boundary. The problem of mining an infinite amount of data is thus converted to mine data that can potentially change the boundary in the current model. Because most of the itemsets do not often change status, which means the boundary is relatively stable, and even if some does, the boundary movement is local, the cost of mining closed frequent itemsets is dramatically reduced.

1.2 Our contribution

This paper makes the following contributions: (1) We introduce a novel algorithm, Moment,¹ to mine closed frequent itemsets over data stream sliding windows. To the best of our knowledge, our algorithm is the first one for mining *closed* frequent itemsets in data streams. (2) We present an in-memory data structure, the *closed enumeration tree* (CET), which monitors closed frequent itemsets as well as itemsets that form the boundary between the closed frequent itemsets and the rest of the itemsets. We show that (i) a status change of any itemset (e.g., from non-frequent to frequent) must come through the boundary itemsets, which means we do not have to monitor itemsets beyond the boundary, and (ii) the boundary is relatively stable, which means the update cost is minimum. (3) We introduce a novel algorithm to maintain the CET in an efficient way. (4) We have done extensive experimental studies to evaluate the performance of the proposed new algorithm. Experiments show that for mining closed frequent itemsets in data streams, Moment has significant performance advantage over representative algorithms for the state-of-the-art approaches.

The rest of the paper is organized as follows. In Sect. 2, we give necessary background in frequent itemset mining. In Sect. 3, we describe in detail our Moment algorithm. In Sect. 4, we have some discussion on the Moment algorithm. In Sect. 5, we introduce related work. In Sect. 6, we give experimental results. We give conclusion in Sect. 7.

2 Problem statement

2.1 Preliminaries

Given a set of items Σ , a database \mathcal{D} wherein each transaction is a subset of Σ , and a threshold f called the *minimum frequency*, $0 < f \leq 1$, the frequent itemset

¹ Maintaining closed frequent itemsets by incremental updates.

mining problem is to find all itemsets that occur in at least $f|\mathcal{D}|$ transactions. For an itemset I , we call the number of transactions in which I occurs the *support* of I . In addition, we define the *minimum support (minsup)* s as $s = f|\mathcal{D}|$.

We assume that there is a lexicographical order among the items in Σ and we use $X \prec Y$ to denote that item X is lexicographically smaller than item Y . Furthermore, an itemset can be represented by a sequence, wherein items are lexicographically ordered. For instance, $\{A, B, C\}$ is represented by ABC , given $A \prec B \prec C$. We also abuse notation by using \prec to denote the lexicographical order between two itemsets. For instance, $AB \prec ABC \prec CD$.

As an example, let $\Sigma = \{A, B, C, D\}$, $\mathcal{D} = \{CD, AB, ABC, ABC\}$, and $s = 2$, then the frequent itemsets are

$$\mathcal{F} = \{(A, 3), (B, 3), (C, 3), (AB, 3), (AC, 2), (BC, 2), (ABC, 2)\}$$

In \mathcal{F} , each frequent itemset is associated with its support in database \mathcal{D} .

2.2 Combinatorial explosion

According to the *a priori* property, any subset of a frequent itemset is also frequent. Thus, algorithms that mine *all* frequent itemsets often suffer from the problem of combinatorial explosion.

Two solutions have been proposed to alleviate this problem. In the first solution (e.g., [4, 10]), only *maximal* frequent itemsets are discovered. A frequent itemset is *maximal* if none of its proper supersets is frequent. The total number of maximal frequent itemsets \mathcal{M} is usually much smaller than that of frequent itemsets \mathcal{F} , and we can derive each frequent itemset from \mathcal{M} . However, \mathcal{M} does not contain information of the support of each frequent itemset unless the itemset is a maximal frequent itemset. Thus, mining only maximal frequent itemsets loses information.

In the second solution (e.g., [20, 21]), only *closed* frequent itemsets are discovered. An itemset is closed if none of its proper supersets has the same support as it has. Usually, the total number of closed frequent itemsets \mathcal{C} is still much smaller than that of frequent itemsets \mathcal{F} . Furthermore, we can derive \mathcal{F} from \mathcal{C} , because a frequent itemset I must be a subset of one (or more) closed frequent itemset, and I 's support is equal to the maximal support of those closed itemsets that contain I .

In summary, the relation among \mathcal{F} , \mathcal{C} , and \mathcal{M} is $\mathcal{M} \subseteq \mathcal{C} \subseteq \mathcal{F}$. The closed and maximal frequent itemsets for the above examples are

$$\begin{aligned} \mathcal{C} &= \{(C, 3), (AB, 3), (ABC, 2)\} \\ \mathcal{M} &= \{(ABC, 2)\} \end{aligned}$$

Since \mathcal{C} is smaller than \mathcal{F} , and \mathcal{C} does not lose information about any frequent itemsets, in this paper, we focus on mining the closed frequent itemsets because they maintain sufficient information to determine all the frequent itemsets as well as their support.

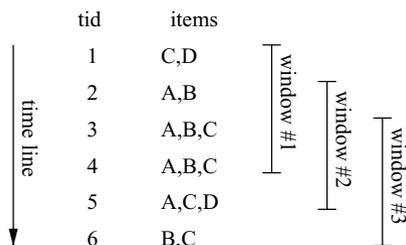


Fig. 1 A running example

2.3 Problem statement

The problem is to mine closed frequent itemsets in the most recent N transactions (or the most recent N *samples*) in a data stream. Each transaction has a time stamp, which is used as the *tid* (transaction id) of the transaction. Figure 1 is an example with $\Sigma = \{A, B, C, D\}$ and window size $N = 4$. We use this example throughout the paper with minimum support $s = 2$.

To find frequent itemsets on a data stream, we maintain a data structure that models the current frequent itemsets. We update the data structure incrementally. The combinatorial explosion problem of mining frequent itemsets becomes even more serious in the streaming environment. As a result, on the one hand, we cannot afford keeping track of all itemsets or even all frequent itemsets, because of time and space constraints. On the other hand, any omission (for instance, maintaining only \mathcal{M} , \mathcal{C} , or \mathcal{F} instead of all itemsets) may prevent us from discovering future frequent itemsets. Thus, the challenge lies in designing a compact data structure which does not lose information of any frequent itemset over a sliding window.

3 The Moment algorithm

We propose the Moment algorithm and an in-memory data structure, the *closed enumeration tree*, to monitor a dynamically selected small set of itemsets that enable us to answer the query “what are the current closed frequent itemsets?” at any time.

3.1 The closed enumeration tree

Similar to a prefix tree, each node n_I in a *closed enumeration tree* (CET) represents an itemset I . A child node, n_J , is obtained by adding a new item to I such that $I < J$. However, unlike a prefix tree, which maintains *all* itemsets, a CET only maintains a *dynamically selected* set of itemsets, which include (i) closed frequent itemsets, and (ii) itemsets that form a *boundary* between closed frequent itemsets and the rest of the itemsets.

As long as the window size is reasonably large, and the concept drifts in the stream are not too dramatic, most itemsets do not change their status (from frequent to non-frequent or from non-frequent to frequent). In other words, the effects

of transactions moving in and out of a window offset each other and usually do not cause change of status of many involved nodes.

If an itemset does not change its status, nothing needs to be done except for increasing or decreasing the counts of the involved itemsets. If it does change its status, then, as we will show, the change must come through the boundary nodes, which means the changes to the entire tree structure is still limited.

We further divide itemsets on the boundary into two categories, which correspond to the boundary between frequent and non-frequent itemsets, and the boundary between closed and non-closed itemsets, respectively. Itemsets within the boundary also have two categories, namely the closed nodes, and other intermediary nodes that have closed nodes as descendants. For each category, we define specific actions to be taken in order to maintain a shifting boundary when there are concept drifts in data streams (Sect. 3.3). The four types of itemsets are listed below.

1. *Infrequent gateway nodes* A node n_I is an infrequent gateway node if (i) I is an infrequent itemset, (ii) n_I 's parent, n_J , is frequent, and (iii) I is the result of joining I 's parent, J , with one of J 's frequent siblings. In addition, we define all nodes at the first level of the CET tree that correspond to infrequent items as infrequent gateway nodes. In Fig. 2, D is an infrequent gateway node (represented by dashed circle). In contrast, AD is not an infrequent gateway node (hence it does not appear in the CET), because D is infrequent.
2. *Unpromising gateway nodes* A node n_I is an unpromising gateway node if (i) I is a frequent itemset, and (ii) there exists a closed frequent itemset J such that $J \prec I$, $J \supset I$, and J has the same support as I does. In Fig. 2, B is an unpromising gateway node because AB has the same support as B does. So is AC because of ABC . In Fig. 2, unpromising gateway nodes are represented by dashed rectangles. For convenience of discussion, when a node in the CET is neither an infrequent gateway node nor an unpromising gateway node, we call it a *promising node*.
3. *Intermediate nodes* A node n_I is an intermediate node if (i) I is a frequent itemset, (ii) n_I has a child node n_J such that J has the same support as I does, and (iii) n_I is not an unpromising gateway node. In Fig. 2, A is an intermediate node because its child AB has the same support as A does.
4. *Closed nodes* These nodes represent closed frequent itemsets in the current sliding window. A closed node can be an internal node or a leaf node. In Fig. 2, C , AB , and ABC are closed nodes, which are represented by solid rectangles.

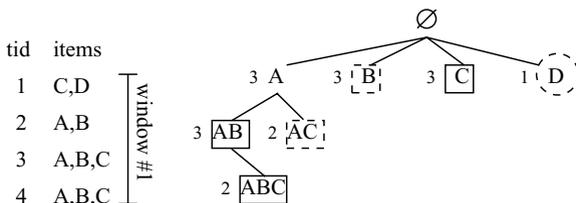


Fig. 2 The closed enumeration tree corresponding to window #1 (each node is labeled with its support)

3.2 Node properties

We prove the following properties for the nodes in the CET. Properties 1 and 2 enable us to prune a large amount of itemsets from the CET, while Property 3 makes sure certain itemsets are not pruned. Together, they enable us to mine closed frequent itemsets over a sliding window using an efficient and compact synopsis data structure.

Property 1 If n_I is an infrequent gateway node, then any node n_J where $J \supset I$ represents an infrequent itemset.

Proof Property 1 is derived from the *a priori* property. □

A CET achieves its compactness by pruning a large amount of the itemsets. It prunes the descendants of n_I and the descendants of n_I 's siblings nodes that subsume I . However, the CET 'remembers' the boundary where such pruning occurs, so that it knows where to start exploring when n_I is no longer an infrequent gateway node. An infrequent gateway node marks such a boundary. In particular, infrequent gateway nodes are leaf nodes in a CET. For example, in Fig. 2, after knowing that D is infrequent, we do not explore the subtree under D . Furthermore, we do not join A with D to generate A 's child nodes. As a result, a large amount of the itemsets are pruned.

Property 2 If n_I is an unpromising gateway node, then n_I is not closed, and none of n_I 's descendants is closed.

Proof Based on the definition of unpromising gateway nodes, there exists an itemset J such that (i) $J \prec I$, and (ii) $J \supset I$ and $support(J) = support(I)$. From (ii), we know n_I is not closed. Let i_{max} be the lexicographically largest item in I . Since $J \prec I$ and $J \supset I$, there must exist an item $j \in J \setminus I$ such that $j \prec i_{max}$. Thus, for any descendant $n_{I'}$ of n_I , we have $j \notin I'$. Furthermore, because $support(J) = support(I)$, itemset $J \setminus I$ must appear in every transaction I appears, which means $support(n_{I'}) = support(n_{\{j\} \cup I'})$, so I' is not closed. □

Descendants of an unpromising gateway node are pruned because no closed nodes can be found there, and the CET 'remembers' the boundary where such pruning occurs by recording the unpromising gateway nodes.

Property 3 If n_I is an intermediate node, then n_I is not closed and n_I has closed descendants.

Proof Based on the definition of intermediate nodes, n_I is not closed. Thus, there must exist a closed node n_J such that $J \supset I$ and $support(J) = support(I)$. If $J \prec I$, then n_I is an unpromising gateway node, which means n_I cannot be an intermediate node. So we have $I \prec J$. However, if $I \prec J$, then n_J must be n_I 's descendant because $J \supset I$. □

Property 3 shows that we cannot prune intermediate nodes in a CET.

3.3 Building the closed enumeration tree

For each node n_I in a CET, we store the following information: (i) the itemset I itself,² (ii) the node type of n_I , (iii) *support*: the number of transactions in which I occurs, and (iv) *tid_sum*: the sum of the tids of the transactions in which I occurs. The purpose of having *tid_sum* is because we use a hash table to maintain closed itemsets.

3.3.1 The hash table

We frequently check whether or not a certain node is an unpromising gateway node, which means we need to know whether there is a closed frequent node that has the same support as the current node.

We use a hash table to store all the closed frequent itemsets. To check if n_I is an unpromising gateway node, by definition, we check if there is a closed frequent itemset J such that $J \prec I$, $J \supset I$, and $\text{support}(J) = \text{support}(I)$.

We can thus use *support* as the key to the hash table. However, it may create frequent hash collisions. We know if $\text{support}(I) = \text{support}(J)$ and $I \subset J$, then I and J must occur in the same set of transactions. Thus, a better choice is the set of *tids*. However, the set of *tids* take too much space, so we instead use $(\text{support}, \text{tid_sum})$ as the key. Note that *tid_sum* of an itemset can be incrementally updated. To check if n_I is an unpromising gateway node, we hash on the $(\text{support}, \text{tid_sum})$ of n_I , fetch the list of closed frequent itemsets in the corresponding entry of the hash table, and check if there is a J in the list such that $J \prec I$, $J \supset I$, and $\text{support}(J) = \text{support}(I)$.

To save space, in the hash table entries, instead of the itemsets themselves, we store the pointers pointing to the corresponding nodes in the CET.

3.3.2 FP-tree for transactions

We store the transactions in the sliding window in an FP-tree, in order to reduce the memory footprint and to speed up exploration of the transactions. FP-tree was first introduced by Han et al. for mining frequent itemsets without candidate generation [11]. In an FP-tree, each transaction is stored along a root-path; when transactions have a common prefix, the common part only needs to be stored once; a counter is used to record the number of times the common part is repeated. As demonstrated by Han et al., an FP-tree is a compact data structure that stores all necessary information for frequent itemsets mining and it is usually much smaller than the database itself. Figure 3 shows the FP-tree for the first sliding window. Note that the items are stored in an inverse lexicographical order among the root-path. This arrangement makes it easy to explore the FP-tree.

Our FP-tree is slightly different from the one we described above. First, we use the FP-tree to store all the transactions in the sliding window, so we do not prune infrequent items. Second, in addition to the head table in traditional FP-trees (which is used to record the starting pointers to each item), we also maintain

² In our implementation, we do not actually store the whole itemset I in node n_I – instead, we only store the last item in I . Because we always visit a node following a root-path of the CET, we can derive the itemset I by concatenating the items stored in the nodes along the root-path.

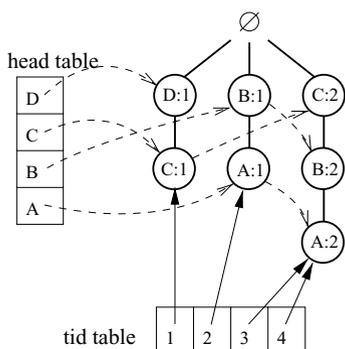


Fig. 3 The FP-tree for transactions in the sliding window

another table, the tid table. In the tid table, for each tid (transaction id), there is a pointer pointing to a node in the FP-tree, which we call the node the *tail* of the transaction; the path from the tail to the root of the FP-tree gives us the itemset corresponding to the given tid. By using the FP-tree with the tid table, we do not need the transactions anymore.

To add a transaction to the sliding window, we store the corresponding itemset in the FP-tree and insert a new entry at the end of the tid table, where the pointer in the new entry points to the tail of the new transaction in the FP-tree; to delete a transaction from the sliding window, we pop an entry from the front of the tid table, and use the pointer to locate in the FP-tree the tail of the transaction to be deleted. We then follow the path from the tail to the root of the FP-tree, and update the counters along the path. Notice that although the size of the tid table is the same as that of the sliding window (N), if we follow a first-in-first-out rule for updating the sliding window, most part of the tid table can be stored in disk, because we only update the front and the end of the tid table.

3.3.3 CET construction

To build a CET, first we create a root node n_{\emptyset} . Second, we create $|\Sigma|$ child nodes for n_{\emptyset} (i.e., each $i \in \Sigma$ corresponds to a child node $n_{\{i\}}$), and then we call *Explore* on each child node $n_{\{i\}}$. Pseudo code for the *Explore* algorithm is given in Fig. 4.

Explore: This is a depth-first procedure that visits itemsets in lexicographical order. For an itemset I , *Explore* consults the FP-tree to determine the *support* and *tid_sum* of I . In lines 1 and 2 of Fig. 4, if a node is found to be infrequent, then it is marked as an infrequent gateway node, and we do not explore it further (Property 1). However, the *support* and *tid_sum* of an infrequent gateway node have to be stored because they will provide important information during a CET update when an infrequent itemset can potentially become frequent.

In lines 3 and 4, when an itemset I is found to be non-closed because of another lexicographically smaller itemset, then n_I is an unpromising gateway node. Based on Property 2, we do not explore n_I 's descendants, which does not contain any closed frequent itemsets. However, n_I 's *support* and *tid_sum* must be stored, because during a CET update, n_I may become promising.

```

Explore ( $n_I, \mathcal{D}, \text{minsup}$ )
1: if  $\text{support}(n_I) < \text{minsup}$  then
2:   mark  $n_I$  an infrequent gateway node;
3: else if  $\text{leftcheck}(n_I) = \text{true}$  then
4:   mark  $n_I$  an unpromising gateway node;
5: else
6:   foreach frequent right sibling  $n_K$  of  $n_I$  do
7:     create a new child  $n_{I \cup K}$  for  $n_I$ ;
8:     compute  $\text{support}$  and  $\text{tid\_sum}$  for  $n_{I \cup K}$ ;
9:   foreach child  $n_{I'}$  of  $n_I$  do
10:     $\text{Explore}(n_{I'}, \mathcal{D}, \text{minsup})$ ;
11:  if  $\exists$  a child  $n_{I'}$  of  $n_I$  such that
     $\text{support}(n_{I'}) = \text{support}(n_I)$  then
12:    mark  $n_I$  an intermediate node;
13:  else
14:    mark  $n_I$  a closed node;
15:    insert  $n_I$  into the hash table;

```

Fig. 4 The *Explore* algorithm

In *Explore*, $\text{leftcheck}(n_I)$ checks if n_I is an unpromising gateway node. It looks up the hash table to see if there exists a previously discovered closed itemset that has the same support as n_I and which also subsumes I , and if so, it returns *true* (in this case n_I is an unpromising gateway node); otherwise, it returns *false* (in this case n_I is a promising node).

If a node n_I is found to be neither infrequent nor unpromising, then we explore its descendants (lines 6–10). After that, we can determine if n_I is an intermediate node or a closed node (lines 11–15) according to Property 3.

Complexity: The time complexity of the *Explore* algorithm depends on the size of the sliding window N , the minimum support, and the number of nodes in the CET. However, because *Explore* only visits those nodes that are necessary for discovering closed frequent itemsets, so *Explore* should have the same asymptotic time complexity as any closed frequent itemset mining algorithm that is based on traversing the enumeration tree.

3.4 Updating the CET

New transactions are inserted into the window, as old transactions are deleted from the window. We discuss the maintenance of the CET for the two operations: addition and deletion.³

3.4.1 Adding a transaction

In Fig. 5, a new transaction T (*tid* 5) is added to the sliding window. We traverse the parts of the CET that are related to transaction T . For each related node n_I , we update its *support*, *tid_sum*, and possibly its node type.

³ At the time that a new transaction is added to the sliding window, the window size is temporarily increased to $N + 1$; after that, deleting a transaction from the sliding window will change the window size back to N . Therefore in our algorithm, we assume that the minimum support (*minsup*) remained unchanged during the addition and the deletion.

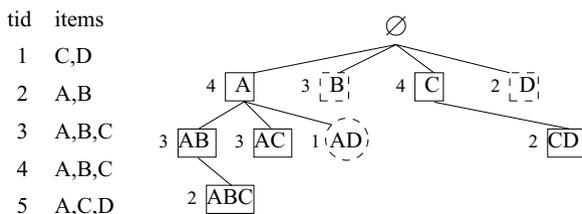


Fig. 5 Adding a transaction

Most likely, n_I 's node type will not change, in which case, we simply update n_I 's support and tid_sum , and the cost is minimum. In the following, we discuss cases where the new transaction T causes n_I to change its node type.

n_I was an infrequent gateway node. If n_I becomes frequent (e.g., from node D in Fig. 2 to node D in Fig. 5), two types of updates must be made. First, for each of n_I 's left siblings it must be checked if new children should be created. Second, the originally pruned branch (under n_I) must be re-explored by calling *Explore*.

For example, in Fig. 5, after D changes from an infrequent gateway node to a frequent node, node A and C must be updated by adding new children (AD and CD , respectively). Some of these new children will become new infrequent gateway nodes (e.g., node AD), and others may become other types of nodes (e.g., node CD becomes a closed node). In addition, this update may propagate down more than one level.

n_I was an unpromising gateway node. Node n_I may become promising (e.g., from node AC in Fig. 2 to node AC in Fig. 5) for the following reason. Originally, $\exists(j < i_{max} \text{ and } j \notin I)$ s.t. j occurs in each transaction that I occurs. However, if T contains I but not any of such j 's, then the above condition does not hold anymore. If this happens, the originally pruned branch (under n_I) must be explored by calling *Explore*.

n_I was a closed node. Based on the following property, n_I will remain a closed node.

Property 4 Adding a new transaction will not change a node from closed to non-closed, and therefore it will not decrease the number of closed itemsets in the sliding window.

Proof Originally, $\forall J \supset I, support(J) < support(I)$; after adding the new transaction T , $\forall J \supset I$, if $J \subset T$ then $I \subset T$. Therefore if J 's support is increased by one because of T , so is I 's support. As a result, $\forall J \supset I, support(J) < support(I)$ still holds after adding the new transaction T . However, if a closed node n_I is visited during an addition, its entry in the hash table will be updated. Its support is increased by 1 and its tid_sum is increased by adding the tid of the new transaction. □

n_I was an intermediate node. An intermediate node, such as node A in Fig. 2, can possibly become a closed node after adding a new transaction T . Originally, n_I was an intermediate node because one of n_I 's children has the same support as n_I does; if T contains I but none of n_I 's children who have the same support as n_I

Addition ($n_I, I_{new}, \mathcal{D}, minsup$)

```

1: if  $n_I$  is not relevant to the addition then return;
2: foreach child node  $n_{I'}$  of  $n_I$  do
3:   update support and tid_sum of  $n_{I'}$ ;
4:  $\mathcal{F} \leftarrow \{n_{I'} | n_{I'} \text{ is newly frequent}\}$ ;
5: foreach child node  $n_{I'}$  of  $n_I$  do
6:   if  $n_{I'}$  is infrequent then
7:     (re)mark  $n_{I'}$  an infrequent gateway node;
8:   else if leftcheck( $n_{I'}$ ) = true then
9:     (re)mark  $n_{I'}$  an unpromising gateway node;
10:  else if  $n_{I'}$  is a newly frequent node or
       $n_{I'}$  is a newly promising node then
11:    Explore( $n_{I'}$ ,  $\mathcal{D}$ , minsup);
12:  else
13:    foreach  $n_K \in \mathcal{F}$  s.t.  $I' \prec K$  do
14:      add  $n_{I' \cup K}$  as a new child of  $n_{I'}$ ;
15:      Addition( $n_{I'}$ ,  $I_{new}$ ,  $\mathcal{D}$ , minsup);
16:    if  $n_{I'}$  was a closed node then
17:      update  $n_{I'}$ 's entry in the hash table;
18:    else if  $\exists$  a child node  $n_{I''}$  of  $n_{I'}$  s.t.
      support( $n_{I''}$ ) = support( $n_{I'}$ ) then
19:      mark  $n_{I'}$  a closed node;
20:      insert  $n_{I'}$  into the hash table;
21: return;
```

Fig. 6 The *Addition* algorithm

had before the addition, then n_I becomes a closed node because its new support is higher than the support of any of its children. However, n_I cannot change to an infrequent gateway node or an unpromising gateway node. First, n_I 's support will not decrease because of adding T , so it cannot become infrequent. Second, if before adding T , *leftcheck*(n_I) = false, then $\exists(j < i_{\max} \text{ and } j \notin I)$ s.t. j occurs in each transaction that I occurs; this statement will not change after we add T . Therefore, *leftcheck*(n_I) = false after the addition.

Figure 6 gives a high-level description of the addition operation. Adding a new transaction to the sliding window will trigger a call of *Addition* on n_\emptyset , the root of the CET.

From the above discussion and from the *Addition* algorithm shown in Fig. 6, we can easily derive the following property of *Addition*:

Property 5 The *Addition* algorithm will not decrease the number of nodes in a CET.

3.4.2 Deleting a transaction

In Fig. 7, an old transaction T (*tid* 1) is deleted from the sliding window. To delete a transaction, we also traverse the parts of the CET that is related to the deleted transaction. Most likely, n_I 's node type will not change, in which case, we simply update n_I 's *support* and *tid_sum*, and the cost is minimum. In the following, we discuss the impact of deletion in detail.

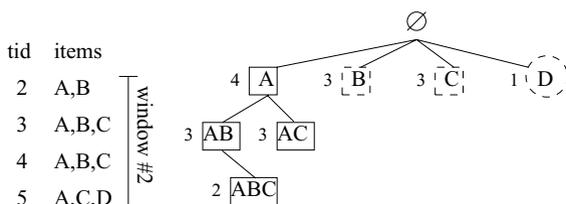


Fig. 7 Deleting a transaction

If n_I was an infrequent gateway node, obviously deletion does not change n_I 's node type. If n_I was an unpromising gateway node, deletion may change n_I to infrequent but will not change n_I to promising, for the following reason. For an unpromising gateway node n_I , if before deletion, $leftcheck(n_I) = true$, then $\exists(j < i_{max} \text{ and } j \notin I)$ s.t. j occurs in each transaction that I occurs; this statement remains true when we delete a transaction.

If n_I was a frequent node, it may become infrequent because of a decrement of its support, in which case, all n_I 's descendants are pruned and n_I becomes an infrequent gateway node. In addition, all of n_I 's left siblings are updated by removing children obtained from joining with n_I . For example in Fig. 7, when transaction T (tid 1) is removed from the window, node D becomes infrequent. We prune all descendants of node D , as well as AD and CD , which were obtained by joining A and C with D , respectively.

If n_I was a promising node, it may become unpromising because of the deletion, for the following reason. If before the deletion, $\exists(j < i_{max} \text{ and } j \notin I)$ s.t. j occurs in each transaction that I occurs, except only for the transaction to be deleted, then after deleting the transaction, I becomes unpromising. This happens to node C in Fig. 7. Therefore, if originally n_I was neither infrequent nor unpromising, then we have to do the $leftcheck$ on n_I . From the above discussion we can also see that for a node n_I to change to unpromising because of a deletion, n_I must be contained in the deleted transaction. Therefore n_I will be visited by the traversal and we will not miss it.

If n_I was a closed node, it may become non-closed. To demonstrate this, we delete another transaction T (tid 2) from the sliding window. Figure 8 shows this example where previously closed node n_I (e.g. A and AB) become non-closed because of the deletion. This can be determined by looking at the supports of the children of n_I after visiting them. If a previously closed node that is included in the deleted transaction remains closed after the deletion, we still need to update its entry in the hash table: its $support$ is decreased by 1 and its tid_sum is decreased by subtracting the tid of the deleted transaction.

From the above discussion we derive the following property for the deletion operation on a CET.

Property 6 Deleting an old transaction will not change a node in the CET from non-closed to closed, and therefore it will not increase the number of closed itemsets in the sliding window.

Proof If an itemset I was originally non-closed, then before the deletion, $\exists j \notin I$ s.t. j occurs in each transaction that I occurs. Obviously, this fact will not be

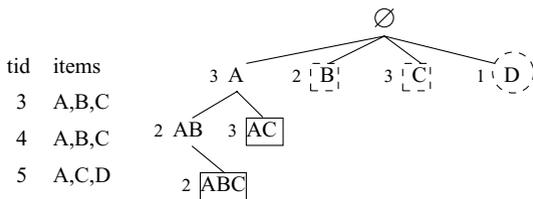


Fig. 8 Another deletion

```

Deletion ( $n_I, I_{old}, minsup$ )
1: if  $n_I$  is not relevant to the deletion then return;
2: foreach child node  $n_{I'}$  of  $n_I$  do
3:   update support and tid.sum of  $n_{I'}$ ;
4:    $\mathcal{F} \leftarrow \{n_{I'} | n_{I'} \text{ is newly infrequent}\}$ ;
5:   foreach child node  $n_{I''}$  of  $n_{I'}$  do
6:     if  $n_{I''}$  was infrequent or unpromising then
7:       continue;
8:     else if  $n_{I''}$  is newly infrequent then
9:       prune  $n_{I''}$ 's descendants from CET;
10:      mark  $n_{I''}$  an infrequent gateway node;
11:    else if leftcheck( $n_{I''}$ ) = true then
12:      prune  $n_{I''}$ 's descendants from CET;
13:      mark  $n_{I''}$  an unpromising gateway node;
14:    else
15:      foreach  $n_K \in \mathcal{F}$  s.t.  $I' \prec K$  do
16:        prune  $n_{I' \cup K}$  from the children of  $n_{I'}$ ;
17:        Deletion( $n_{I'}, I_{old}, minsup$ );
18:      if  $n_{I'}$  was closed and  $\exists$  a child  $n_{I''}$  of  $n_{I'}$ 
19:        s.t. support( $n_{I''}$ ) = support( $n_{I'}$ ) then
20:          mark  $n_{I'}$  an intermediate node;
21:          remove  $n_{I'}$  from the hash table;
21:      else if  $n_{I'}$  was a closed node then
22:        update  $n_{I'}$ 's entry in the hash table;
23:    return;

```

Fig. 9 The Deletion algorithm

changed due to deleting a transaction. So I will still be non-closed after the deletion. □

Figure 9 gives a high-level description of the deletion operation. Some details are skipped in the description. For example, when pruning a branch from the CET, all the closed frequent itemsets in the branch should be removed from the hash table.

From the above discussion and from the *Deletion* algorithm shown in Fig. 9, we can easily derive the following property of *Deletion*:

Property 7 The Deletion algorithm will not increase the number of nodes in a CET.

4 Discussion on the Moment algorithm

In this section, we discuss some properties and extensions of the Moment algorithm.

4.1 Discussion on CET updates

In the addition algorithm, *Explore* is the most time consuming operation, because it scans the transactions stored in the FP-tree. However, as will be demonstrated in the experiments, the number of such invocations is very small, as most insertions will not change node types. In addition, the new branches grown by calling *Explore* are usually very small subsets of the whole CET, therefore such incremental growing takes much less time than regenerating the whole CET. On the other hand, deletion only involves related nodes in the CET, and does not scan transactions stored in the FP-tree. Therefore, its time complexity is at most linear to the number of nodes. Usually it is faster to perform a deletion than an addition.

It is easy to show that if a node n_I changes node type (frequent/infrequent and promising/unpromising), then I is in the added or deleted transaction and therefore n_I is guaranteed to be visited during the update. Consequently, our algorithm will correctly maintain the current close frequent itemsets after any of the two operations. Furthermore, if n_I remains closed after an addition or a deletion and I is contained in the added/deleted transaction, then its position in the hash table is changed because its *support* and *tid_sum* are changed. To make the update, we delete the itemset from the hash table and re-insert it back to the hash table based on the new key value. However, such an update has amortized constant time complexity.

4.2 Variable sliding window size

In our discussion so far, we used sliding windows of fixed size. However, the two operations – *addition* and *deletion* – are independent of each other. Therefore, if needed, the size for the sliding window can grow or shrink without affecting the correctness of our algorithm. However, there is a subtle issue when the size of the sliding window is not fixed. For the discussion so far, we have used an *absolute support*, which is the number of transactions in which an itemset occurs. In most applications, *relative support*, which is the fraction of transactions in which an itemset occurs (i.e., the f in Sect. 2.1), is more commonly used. We have chosen to use absolute support in the Moment algorithm for a practical reason: when updating the CET, instead of all the CET nodes, only those nodes related to the added or deleted transactions are visited, and therefore, Moment can achieve quick response time. When the size of the sliding window is fixed, the Moment algorithm works for both the absolute and relative supports because they are equivalent. However, when the size of the sliding window varies, if we chose to use the relative support, Moment will lose its advantage because an itemset may change its status (in terms of frequent/infrequent and closed/not closed) even if it is not contained in the added or deleted transaction. How to handle relative support under variable sliding window size is among our future work on improving the Moment algorithm.

4.3 Approximate algorithms

The Moment algorithm is deterministic. However, our algorithm does not restrict a deletion to happen at the end of the window: at any given time, any transaction in the sliding window can be removed. As a result, with minor changes to the Moment algorithm, we can implement different approximate algorithms. For one example, when a new transaction arrives, we can use random sampling with a reservoir [19] to decide whether to insert the new transaction into the sliding window and if so, which old transaction to remove from the sliding window. As a consequence, the sliding window always contains a set of uniformly selected samples from the data seen so far. As another example, if when removing a transaction, the transaction to be removed is picked with the following random scheme: the newer transactions have lower probability of being removed than the older ones, then our algorithm can implement a sliding window with *soft* boundary, i.e., the more recent the transaction, the higher chance it will remain in the sliding window.

4.4 Streaming output

From the description of the Moment algorithm (Figs. 6 and 9) we can see that when adding or deleting a transaction, only those CET nodes related to the update will be visited and changed. One consequence of this design, as we mentioned before, is the efficiency of the algorithm. In addition, another consequence is that the algorithm can directly output those itemsets that have changed status due to the update. For example, the Moment algorithm can output the itemsets that have become newly closed frequent and the itemsets that have ceased to have that property due to the most recent update. These itemsets can be output as data streams and be sent to other data analysis processes.

4.5 Lazy pruning

In the Moment algorithm, to reduce memory usage, all descendants of an infrequent gateway node or an unpromising gateway node are pruned from the CET. In some cases, however, this design is not time efficient. For example, if the newly added transaction is exactly the same as the just deleted transaction, then all the CET nodes pruned due to the previous deletion must be regrown because of the addition, and we know growing nodes (by calling *explor()*) is time consuming. A possible method to alleviate this situation is to use lazy pruning. That is, instead of physically pruning a node from CET, we prune it logically by using a flag to indicate that the node is no longer a part of the CET. If this node has to be regrown later, we can simply change the flag instead of calling *explor()*. Of course, such a solution has its own problems. On the one hand, because of combinatoric explosion, it is impossible to maintain all infrequent nodes, and therefore we have to make a decision on which nodes are to be physically pruned and which nodes are to be logically pruned. On the other hand, for a logically pruned node to be useful in the future, the information on its support must be maintained. That is, during an update, in addition to the CET nodes that are related to the added/deleted transaction, we have to visit the logically deleted nodes related to the transaction to

update their supports. Currently, we are investigating on how to add this extension to the Moment algorithm.

4.6 Batch updates

So far our algorithm only handles one transaction in one update. There are situations in which data are bursty and multiple transactions need to be added and deleted during one update. However, it is not difficult to adapt our algorithm to handle multiple transactions in one update. Originally, for an addition or a deletion, we traverse the CET with the single added or deleted transaction; if an update contains a batch of transactions, we can still traverse the CET in the same fashion using the batch of transactions and project out unrelated transactions along the traversal.

5 Related work

Incrementally mining frequent itemsets has been investigated by many researchers. Manku et al. [13] developed a randomized algorithm, the Sticky Sampling Algorithm, and a deterministic algorithm, the Lossy Counting Algorithm, for maintaining frequent *items* over a data stream where for a given time t , the frequent items are defined over the *entire* data stream up to t . The algorithms guarantee no false negative and a bound on the error of estimated frequency (the guarantees are in a probabilistic sense for the randomized algorithm). The Lossy Counting Algorithm is extended to handle frequent *itemsets*, where a trie is used to maintain all frequent itemsets and the trie is updated by batches of transactions in the data stream. The main differences between Manku's algorithms and our Moment algorithm are (1) in Manku's algorithms, the frequent items (and itemsets) are defined over the whole data stream while in Moment the frequent itemsets are defined over a sliding window to reflect the most recent trends, (2) Manku's algorithms are approximate algorithms where the support of an itemset is only guaranteed to be within a range while in Moment, the support is exact, and (3) Manku's algorithms strive for a tunable compromise between memory usage and error bounds while in Moment, quick response time is the main goal.

Cheung et al. [7, 8] proposed algorithms *FUP* and *FUP*₂ for incrementally updating frequent itemsets. Thomas et al. [17] presented a similar algorithm. Both Cheung's and Thomas's algorithms assume batch updates and take advantage of the relationship between the original database (*DB*) and the incrementally changed transactions (*db*). *FUP* is similar to the well-known Apriori Algorithm, which is a multiple-step algorithm. The key observation of *FUP* is that by adding *db* to *DB*, some previously frequent itemsets will remain frequent and some previously infrequent itemsets will become frequent (these itemsets are called *winners*); at the same time, some previously frequent itemsets will become infrequent (these itemsets are called *losers*). The key technique of *FUP* is to use information in *db* to filter out some winners and losers, and therefore reduce the size of candidate set in the Apriori algorithm. Because the performance of the Apriori algorithm relies heavily on the size of candidate set, *FUP* improves the performance of Apriori greatly. *FUP*₂ extended *FUP* by allowing deleting old transactions from a database

as well. The algorithm proposed by Thomas et al. is similar to FUP_2 except that in addition to frequent itemsets, a negative border [14] is maintained. In the algorithm, the frequent itemsets in db are mined first. At the same time, the counts of frequent itemsets (and itemsets on the negative border) in DB are updated. Then based on the change of the frequent itemsets in DB , the negative border in DB , and the frequent itemsets in db , the frequent itemsets in the updated database are computed with a possible scan of the updated database. Because the updated database is scanned at most once, Thomas's algorithm has a very good performance rate. Cheung's and Thomas's algorithms are different from our Moment algorithm in several ways. First, in Cheung's and Thomas's algorithms, the update is assumed to be a batch file db and if the size of db is too small, the algorithm will not work because in such a case, almost all itemsets in db will be frequent (as an example consider the case when $|db| = 2$ and $minsup = 50\%$). We argue that in data stream applications, updates are high speed and users are often interested in the new frequent itemsets in real time. Accumulating a batch file db before running the algorithm will prolong the response time and is not appropriate for data stream applications. Second, in both Cheung's and Thomas's algorithms, for one updates, all frequent itemsets need to be updated: in Cheung's algorithms, although some candidates are pruned by the algorithms, the Apriori algorithm is run once for each update; in Thomas's algorithm, for each update the whole negative border needs to be regenerated from scratch. In contrast, in our Moment algorithm, only the part of the boundary that is related to the change needs to be updated.

Veloso et al. [18] proposed an algorithm ZIGZAG for mining frequent itemsets in evolving databases. Later, Otey et al. [15] extended ZIGZAG into parallel and distributed algorithms. ZIGZAG is similar to Cheung's and Thomas's algorithms in that it achieves its speedup by using the relationship between DB and db . However, ZIGZAG has many distinct features. First, ZIGZAG mainly used db to speedup the support counting of frequent itemsets in the updated database and it does not discover the frequent itemsets in db itself. As a result, for a given minimum support, ZIGZAG can handle batch update with arbitrary block size. Second, ZIGZAG adapts the techniques proposed in the GENMAX algorithm [10] and in each update only maintains maximal frequent itemsets. Because the information on maximal frequent itemsets and their supports is not enough to generate association rules (because the support information of some non-maximal frequent itemsets may be missing), a second step is used in ZIGZAG in which the updated database is scanned to discover all frequent itemsets and their supports. In our experimental study, we will compare the performance of our algorithm with that of ZIGZAG.

Charikar et al. [6] presented a one-pass algorithm that returns most frequent *items* whose frequencies satisfy a threshold with high probabilities. Teng et al. [16] presented an algorithm, FTP-DS, that mines frequent temporal patterns from data streams of itemsets. Chang et al. [5] presented an algorithm, *estDec*, that mines recent frequent itemsets where the frequency is defined by an aging function. Giannella et al. [9] proposed an approximate algorithm for mining frequent itemsets in data streams during arbitrary time intervals. An in-memory data structure, *FP-stream*, is used to store and update historic information about frequent itemsets and their frequency over time and an aging function is used to update the entries so that more recent entries are weighted more. Asai et al. [3] presented an

online algorithm, *StreamT*, for mining frequent rooted ordered trees. To reduce the number of subtrees to be maintained, an update policy that is similar to that in online association rule mining [12] was used and therefore the results are inexact. In all these studies, approximate algorithms were used. In contrast, our algorithm is an exact one. On the other hand, we can also assume that an approximation step has been implemented through the sampling scheme and our exact algorithm works on a sliding window containing the random samples (which are a synopsis of the data stream).

In addition to the differences mentioned earlier, one distinct feature of our algorithm is that it only mines and maintains *closed* frequent itemsets while all above algorithms focused on mining *all* frequent itemsets. The large number of frequent itemsets makes it impractical to maintain information about all frequent itemsets using in-memory data structures. As demonstrated by extensive experimental studies, e.g., [21], there are usually much fewer closed frequent itemsets compared to the total number of frequent itemsets. As a consequence, our algorithm has better performance in terms of memory usage and running time, as is demonstrated in the experimental studies.

6 Experimental results

We performed extensive experiments to evaluate the performance of Moment algorithm. We use Charm, a state-of-the-art algorithm proposed by Zaki et al. [21], as the baseline algorithm to generate closed frequent itemsets without using incremental updates. We have used the latest version of Charm. As demonstrated in many studies (e.g., [21, 22]), among the algorithms that mine closed frequent itemsets, Charm has best performance for various data sets. We also compare the performance of Moment with that of ZIGZAG, a recently proposed algorithm on incrementally mining frequent itemsets [18]. All our experiments were done on a 2 GHz Intel Pentium IV PC with 2 GB main memory, running RedHat Linux 7.3 operating system.

For the performance study, we have used three synthetic data sets and six real-world data sets. The synthetic data sets are generated using the IBM synthetic data generator for association rules. The first three real-world data sets were used for KDDCUP 2000 [22]. Among them, the first two, BMS-WebView-1 and BMS-WebView-2, record several months of clickstream data from two e-commerce web sites; the third one, BMS-POS, contains several years of point-of-sale data from a large electronics retailer [22]. The fourth real-world data set, Mushroom, has been used extensively in the AI area. The last two real-world data sets, WCup and WPortal, are kindly provided by Matthew Eric Otey [15]. Among them, WCup is derived from the click-stream data of the official 1998 World Soccer Cup web site, and WPortal is derived from a large web portal in Brazil. The data characteristics for all the data sets are summarized in Table 1.

6.1 Synthetic data sets

The synthetic data sets are generated using the synthetic data generator developed by Agrawal et al. [2]. Data from this generator mimics transactions from retail stores. Here are some of the parameters that we have controlled: the size of the

Table 1 Data characteristics

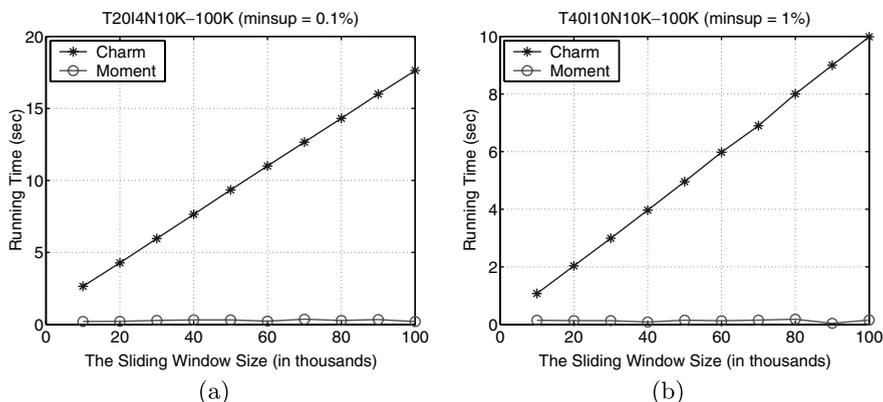
Database	No. of items	Average length	Max length	# Records	Window size
T20I4N10 K–100 K	1,000	20	44	100,000	10 K–100 K
T40I10N10 K–100 K	1,000	40	80	100,000	10 K–100 K
T20I10N100 K	1,000	20	49	100,000	2 K, 100 K
BMS-WebView-1	497	2.5	267	59,602	2 K, 50 K
BMS-WebView-2	3,340	4.6	161	77,512	2 K, 50 K
BMS-POS	1,657	6.5	164	515,597	2 K, 500 K
Mushroom	120	23	23	8,124	2 K, 8 K
WCup	16,788	7.2	100	2,222,000	2 K, 100 K
WPortal	5,864	1.8	89	3,500,000	2 K, 100 K

sliding window N , the average size of transactions T , the average size of the maximal potentially frequent itemsets I .

6.1.1 Performance under different sliding window sizes

In the first experiment, we compare Moment and Charm under different sliding window sizes. For this study, we generated two data sets: in the first one, T20I4N10 K–100 K, we have set the parameters as $T = 20$, $I = 4$; in the second one, T40I10N10 K–100 K, we have set the parameters as $T = 40$, $I = 10$. In both data sets, we let the sliding window size N grow from 10 K to 100 K. Because in Moment, one update consists of adding a new transaction to and deleting an old transaction from a sliding window, each sliding window differs from the previous one by exactly one transaction. That is, for example when the sliding window size is 10,000, the first sliding window contains transactions 1–10,000, the second sliding window contains transactions 2–10,001, and so on. In the experimental results, for both algorithms, we report the average running time over 100 consecutive sliding windows.

As shown in Fig. 10, as the sliding window size increases, the time to generate all closed frequent itemsets for Charm grows in a linear fashion. In contrast, the running time of Moment does not change too much with the sliding window

**Fig. 10** Running time vs. sliding window size

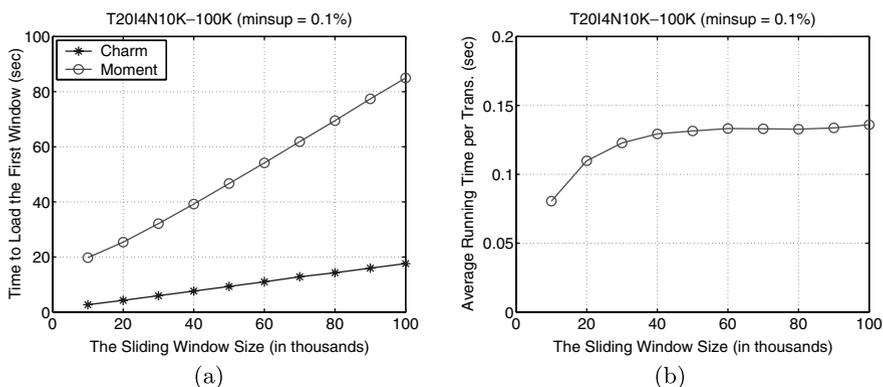


Fig. 11 Bulk loading vs. incremental loading

size. This result demonstrates an advantage of the Moment algorithm: because of its *incremental* updating fashion, it is not sensitive to the sliding window size.

In the above experiments, for the Moment algorithm to start running incrementally, the information about the first sliding window should be available. In Fig. 11a we compare the time for Moment to bulk-load the first sliding window (by calling *Explore()*) and the time for Charm to mine the closed frequent itemsets in the first sliding window. As can be seen from the figure, for getting the results in the first sliding window, Charm is faster by 5–10 times. There are several reasons for this result: first, we have used the latest version of Charm, which is heavily optimized for large set operations (e.g., by using the *diffset* techniques); second, Moment has extra data structures to maintain (e.g., creating the CET nodes, update their support and *tid_sum*, etc.). However, we argue that this comparison is not fair – Moment is an incremental algorithm and the bulk-loading should not be used at all. That is, because data streams are unbounded, there is no such a thing as “the first sliding window”. Of course, there are some special scenarios in which bulk-loading is needed. For example, after a web server had been shut down (e.g., for the purpose of database maintenance), it has just started to accept new customer requests. In such a scenario, we have to accumulate enough transactions to fill in the first sliding window before the algorithm can start working incrementally. In such a case, Moment can actually output current frequent itemsets as soon as new customer requests arrive, even before the first sliding window is completely filled in. To show this point, we have done the following experiment: originally, the sliding window is empty, then transactions are added one by one until the sliding window is full. As new transactions added, Moment outputs the current closed frequent itemsets even though the sliding window is not completely full. We have done this experiment under different sliding window sizes (10–100K), and in Fig. 11b we report the average time for adding each new transaction, where the time includes the time for updating the FP-tree and that for updating the CET. As we can see from the figure, the average time per transaction is very small and it is not very sensitive to the window size.

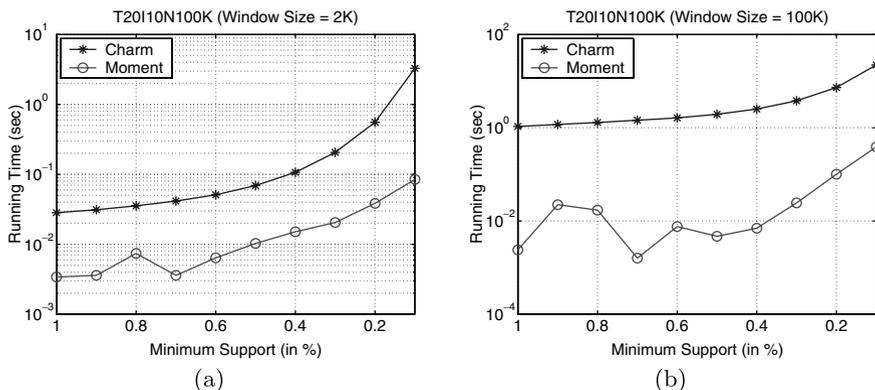


Fig. 12 Performance on T20I10N100 K

6.1.2 Performance under different minimum support

In the second experiment, we compare the performance of Moment and Charm under different minimum supports. The data set we have used, *T20I10N100K*, has the following parameters: $T = 20, I = 10, N = 100K$. We let the minimum support decrease from 1 to 0.1%.

Again, to reduce variation, for each experiment we have executed over 100 consecutive sliding windows and reported the average performance over these 100 sliding windows. Figure 12 shows the average running time for Moment and for Charm under different minimum supports for two different sliding window sizes, 2K and 100K. As can be seen from the figure, as minimum support decreases, because the number of closed frequent itemsets increases, the running time for both algorithms grows. However, the running time of Moment is faster than that of Charm by more than an order of magnitude under all the minimum supports.

Table 2 shows the number of closed itemsets under different minimum supports with the sliding window size of 100K. In addition, in the table we show some static and dynamic statistics about the CET data structure. All reported data are

Table 2 Data characteristics for *T20I10N100 K*

<i>minsup</i> (%)	Closed itemset no.	CET node no.	Changed node no.	Added node no.	Deleted node no.
1.0	721	193,526	0.01	6.20	12.39
0.9	821	210,673	0.01	6.46	0.00
0.8	967	231,376	0.04	13.94	6.76
0.7	1,211	257,498	0.02	0.26	0.00
0.6	1,649	282,330	0.04	0.67	0.27
0.5	2,544	325,834	0.05	1.62	0.16
0.4	4,468	410,629	0.07	6.40	3.17
0.3	9,176	644,622	0.16	5.53	6.67
0.2	22,446	1,549,740	0.59	30.06	48.64
0.1	386,075	7,394,420	38.70	147.67	115.64

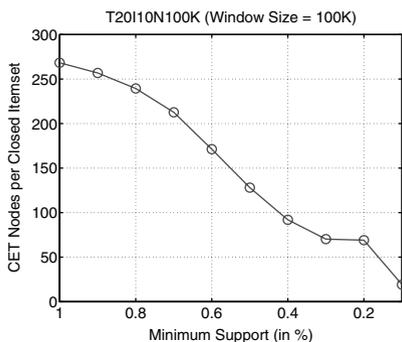


Fig. 13 Performance on T20I10N100 K

average values taken over the 100 sliding windows. The first three columns show the minimum support, the number of closed itemsets, and the number of nodes in the CET. From the table we can see that as the minimum support decreases, the number of closed itemsets grows rapidly. So does the number of nodes in the CET. However, the ratio between the number of nodes in the CET and the number of closed itemsets (which is shown in Fig. 13) actually decreases as the minimum support decreases. This implies that as the sizes of the CET grows larger, it becomes more efficient and the size of the CET is bounded by the number of closed itemsets times a constant number.

Because an addition may trigger a call for *Explore()* which is expensive, we study how many nodes change their status from infrequent/unpromising to frequent/promising (column 4) and how many new nodes are created due to the addition (column 5). From the data we can see that during an addition, the average number of nodes that change from infrequent to frequent or from unpromising to promising in the CET is very small relative to the total number of nodes in the CET. Similarly, the number of new nodes created due to an addition is also very small. These results verify the postulation behind our algorithm: that an update usually only affects the status of a very small portion of the CET and the new branches grown because of an update is usually a very small subset of the CET. In addition, we have reported the average number of CET nodes deleted due to a deletion (column 6). It can be seen that this number is in about the same scale as that of added nodes. However, because a deletion does not query the FP-tree and does not grow the CET, it is a relatively inexpensive operation and therefore will not affect the performance too much.

6.2 Real-world data sets

We have used four real-world data sets to compare the performance of Moment with that of Charm. The first three data sets are BMS-WebView-1, BMS-WebView-2, and BMS-POS. Our fourth real-world data set is the Mushroom data set used by Zaki et al. [21] and it belongs to the family of “dense” data, where there exists strong correlation among transactions. The data characteristics for the 4 data sets are summarized in Table 1.

In the experiments, for each data set we have used two sliding-window sizes, a small one and a large one. We set the small window size to be 2K for all data sets.

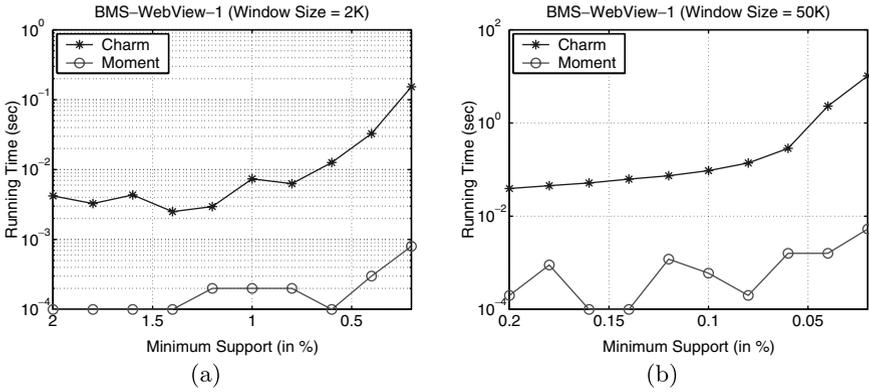


Fig. 14 Performance for BMS-WebView-1

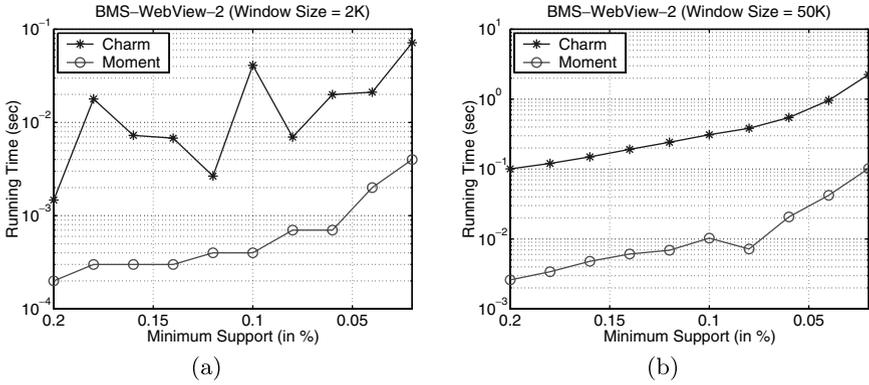


Fig. 15 Performance for BMS-WebView-2

The large window-size is different for different data sets, because the number of available transactions in each data set is different. (Other than these two sliding-window sizes, we have also tested various window sizes in between and got similar results.) Figures 14–17 show the average running time of Moment and Charm for the four data sets, under different minimum supports. From the figure we can see that Moment outperforms Charm by one or two orders of magnitude for all the data sets under a range of values for minimum support and sliding-window size. These results show that Moment has a good performance rate in real-world data sets of various characteristics (sparse or dense data, large or small sliding-window size, large or small minimum support).

6.3 Comparison with ZIGZAG

In this section, we compare the performance of Moment with that of an incremental frequent itemset mining algorithm. Among all the incremental frequent itemset mining algorithms introduced in the section of related work, we were only able to obtain a version of the ZIGZAG algorithm that supports only addition

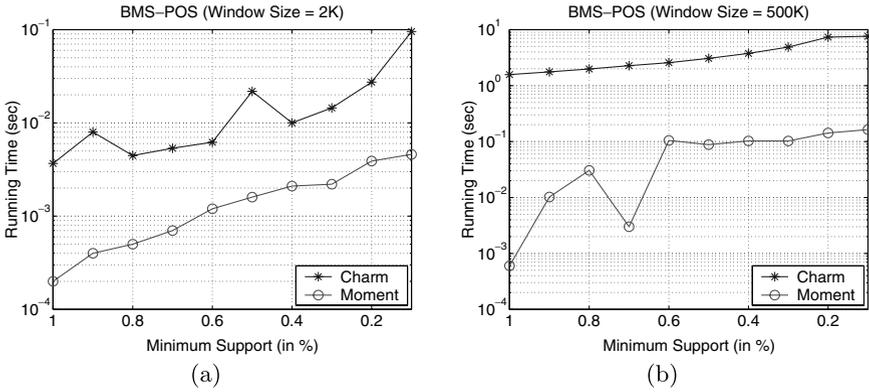


Fig. 16 Performance for BMS-POS

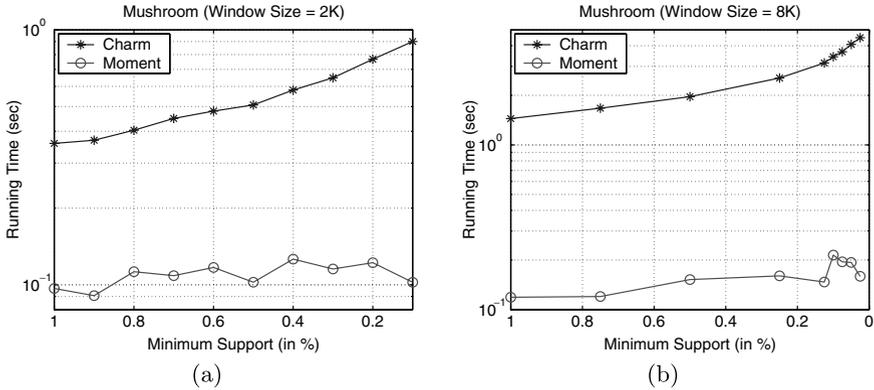


Fig. 17 Performance for Mushroom

(no deletion). Even so, we believe this version of ZIGZAG is representative for the following reasons. First, ZIGZAG is a relatively new algorithm with good performance. For example, in [18], the authors showed that ZIGZAG significantly outperformed Thomas’s algorithm. In addition, Thomas’s algorithm has similar performance to Cheung’s *FUP* and *FUP*₂ algorithms. Second, an update in ZIGZAG can be considered as an addition followed by a deletion. So the running time of the addition-only version of ZIGZAG can be considered as a lower bound of that of the general version of ZIGZAG.

However, incremental algorithms such as ZIGZAG are actually not directly comparable with Moment, because Moment strives for quick response time for updating *one* transaction. Assume for example, that Moment were 100 times faster than a non-incremental frequent itemset mining algorithm in terms of time to update one transaction. In such a case, if users do not mind some delay, they can simply wait until 100 new transactions have been obtained and then apply the non-incremental algorithm to the updated database, which will give them running time similar to that of Moment. Even better, they can apply incremental frequent itemset mining algorithm such as ZIGZAG to get better performance. Therefore,

Table 3 Running-time (in seconds) comparison between Moment and ZIGZAG

Size of batch update	WCup		WPortal	
	Moment	ZIGZAG	Moment	ZIGZAG
Sliding Window Size = 2 K				
1	0.0267	1.24	0.000230	0.0617
10	0.267	1.22	0.00230	0.0618
100	2.67	1.10	0.0230	0.0623
Sliding Window Size = 100 K				
1	0.0980	1.71	0.00139	0.229
10	0.980	1.71	0.0139	0.227
100	9.80	1.71	0.139	0.228

in this section we mainly compare Moment with ZIGZAG in terms of updating one transaction.

For the performance study, we have used the WCup and WPortal data sets provided by Matthew Eric Otey. These two data sets were used by Otey et al. [15] in the performance study of ZIGZAG. The characteristics of the two data sets are described in Table 1. Table 3 gives the running time (in seconds) for Moment and ZIGZAG for the WCup and the WPortal data sets. For each data set, we tested two sliding window sizes, 2K and 100K. For WCup, we have set the minimum support to 0.5%, and for WPortal, 0.1% (for lower minimum support, ZIGZAG will exhaust all available memory). As can be seen from Table 3, for an update consisting of only one transaction (the bold line in the table), Moment outperforms ZIGZAG by orders of magnitude. However, as the batch update size (i.e., the number of transactions in a batch update) grows larger, the running time of ZIGZAG does not change very much. For example, for the WCup data set with the slide window size to be 2K, on average the total time for Moment to make 100 consecutive updates (each consisting of one transaction) is 2.67 s, while the time for ZIGZAG to make one batch update (consisting of 100 transactions) is 1.10 s. These results show that Moment has its limitation: when users are mainly concerned about quick response time for the update of each transaction, Moment is beneficial; however, for a batch update with large batch size, an incremental mining algorithm that can directly handle batch updates is a better choice.

6.4 The number of CET nodes

One design consideration for Moment is to maintain in CET only information related to *closed* frequent itemsets, instead of *all* frequent itemsets. In this section, we use real data sets to justify this consideration.

We show the total number of frequent itemsets, the number of closed frequent itemsets, and the number of CET nodes for two data sets. Figure 18a shows these numbers for the BMS-WebView-1 data set and Fig. 18b shows these numbers for the Mushroom data set under different minimum supports. As can be seen from Fig. 18a, because BMS-WebView-1 is a relatively sparse data set, under high minimum supports, the number of closed frequent itemsets and that of all frequent itemsets do not have much difference; however, when the minimum support decreases further, as some large itemsets become frequent, the total number

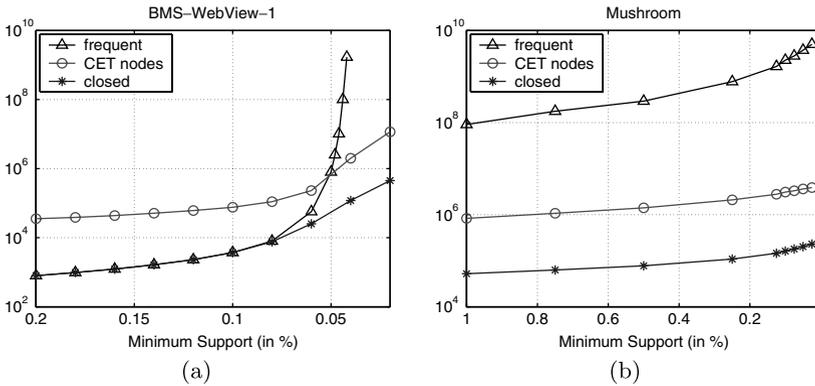


Fig. 18 The number of closed itemsets, CET nodes, and frequent items

of frequent itemsets blows up dramatically; in contrast, the number of CET nodes still keeps a constant ratio relative to the number of closed frequent itemsets. The Mushroom data set, in comparison, is relatively dense, and therefore, even at high minimum support, there are much more frequent itemsets than closed frequent itemsets. As shown in Fig. 18b, although the number of CET nodes is about one order of magnitude more than the number of closed frequent itemsets, it is at least two orders of magnitude fewer than the total number of frequent itemsets.

7 Conclusion

In this paper we propose a novel algorithm, Moment, to discover and maintain all closed frequent itemsets in a sliding window that contains the most recent samples in a data stream. In the Moment algorithm, an efficient in-memory data structure, the closed enumeration tree (CET), is used to record all closed frequent itemsets in the current sliding window. In addition, CET also monitors the items that form the boundary between closed frequent itemsets and the rest of the itemsets. We have also developed efficient algorithms to incrementally update the CET when newly-arrived transactions change the content of the sliding window. Experimental studies show that the running time of the Moment algorithm is not sensitive to the sliding window size and Moment outperforms a state-of-the-art algorithm that mines closed frequent itemsets without using incremental updates. In addition, the number of CET nodes is shown to be proportional to that of closed frequent itemsets. Under low minimum supports or when applied to dense data sets, CET has much fewer number of nodes than the total number of frequent itemsets.

Acknowledgements We thank the anonymous reviewers for their invaluable comments and suggestions. We thank Professor Mohammed J. Zaki for providing the Charm source code. We thank Adriano Alonoso Veloso and Matthew Eric Otey for providing the ZIGZAG source code and offering helpful assistance. The work of these two authors was partly supported by NSF under Grant Nos. 0086116, 0085773, and 9817773. In addition, the work of the first author was partly supported by an IBM summer internship.

References

1. Agarwal RC, Aggarwal CC, Prasad VVV (2001) A tree projection algorithm for generation of frequent item sets. *J Parallel Distrib Comput* 61(3):350–371
2. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large databases (VLDB'94)
3. Asai T, Arimura H, Abe K, Kawasoe S, Arikawa S (2002) Online algorithms for mining semi-structured data stream. In: Proceedings of the 2002 international conference on data mining (ICDM'02)
4. Bayardo Jr RJ (1998) Efficiently mining long patterns from databases. In: Proceedings of the ACM SIGMOD
5. Chang JH, Lee WS (2003) Finding recent frequent itemsets adaptively over online data streams. In: Proceedings of the 2003 international conference knowledge discovery and data mining (SIGKDD'03)
6. Charikar M, Chen K, Farach-Colton M (2002) Finding frequent items in data streams. In: Proceedings of the 29th international colloquium on automata, languages and programming
7. Cheung DW, Han J, Ng V, Wong CY (1996) Maintenance of discovered association rules in large databases: An incremental updating technique. In: Proceedings of the twelfth international conference on data engineering
8. Cheung DW, Lee SD, Kao B (1997) A general incremental technique for maintaining discovered association rules, In: Proceedings of the fifth international conference on database systems for advanced applications (DASFAA)
9. Giannella C, Han J, Robertson E, Liu C (2003) Mining frequent itemsets over arbitrary time intervals in data streams. Technical Report tr587, Indiana University
10. Gouda K, Zaki MJ (2001) Efficiently mining maximal frequent itemsets. In: Proceedings of the 2001 IEEE international conference on data mining
11. Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data
12. Hidber C (1999) Online association rule mining. In: Proceedings of the ACM SIGMOD international conference on management of data
13. Manku G, Motwani R (2002) Approximate frequency counts over data streams. In: Proceedings of the 28th international conference on very large data bases
14. Mannila H, Toivonen H (1997) Levelwise search and borders of theories in knowledge discovery. *Data Mining Knowledge Discov* 1(3):241–258
15. Otey ME, Parthasarathy S, Wang C, Veloso A, Meira W Jr (2004) Parallel and distributed methods for incremental frequent itemset mining. *IEEE Trans Syst Man Cybern, Part B* 34(6):2439–2450
16. Teng W-G, Chen M-S, Yu PS (2003) A regression-based temporal pattern mining scheme for data streams. In: Proceedings of 29th international conference on very large data bases (VLDB'03)
17. Thomas S, Bodagala S, Alsabti K, Ranka S (1997) An efficient algorithm for the incremental updation of association rules in large databases. In: Proceedings of the 1997 international conference knowledge discovery and data mining (SIGKDD'97), pp 263–266
18. Veloso A, Meira Jr W, de Carvalho M, Póssas B, Parthasarathy S, Zaki MJ (2002) Mining frequent itemsets in evolving databases. In: Proceedings of the SDM
19. Vitter JS (1985) Random sampling with a reservoir. *ACM Trans Math Softw* 11(1)
20. Wang J, Han J, Pei J (2003) Closet+: searching for the best strategies for mining frequent closed itemsets. In: Proceedings of the 2003 international conference knowledge discovery and data mining (SIGKDD'03)
21. Zaki MJ, Hsiao C (2002) Charm: an efficient algorithm for closed itemset mining. In: Proceedings of the 2nd SIAM international conference on data mining
22. Zheng Z, Kohavi R, Mason L (2001) Real world performance of association rule algorithms. In: Proceedings of the 2001 international conference on knowledge discovery and data mining (SIGKDD'01)



Yun Chi is currently a Ph.D. student at the Department of Computer Science, UCLA. His main areas of research include database systems, data mining, and bioinformatics. For data mining, he is interested in mining labeled trees and graphs, mining data streams, and mining data with uncertainty.



Haixun Wang is currently a research staff member at IBM T. J. Watson Research Center. He received the B.S. and the M.S. degree, both in computer science, from Shanghai Jiao Tong University in 1994 and 1996. He received the Ph.D. degree in computer science from the University of California, Los Angeles in 2000. He has published more than 60 research papers in referred international journals and conference proceedings. He is a member of the ACM, the ACM SIGMOD, the ACM SIGKDD, and the IEEE Computer Society. He has served in program committees of international conferences and workshops, and has been a reviewer for some leading academic journals in the database field.



Philip S. Yu received the B.S. Degree in electrical engineering from National Taiwan University, the M.S. and Ph.D. degrees in electrical engineering from Stanford University, and the M.B.A. degree from New York University. He is with the IBM Thomas J. Watson Research Center and currently manager of the Software Tools and Techniques group. His research interests include data mining, Internet applications and technologies, database systems, multimedia systems, parallel and distributed processing, and performance modeling. Dr. Yu has published more than 430 papers in refereed journals and conferences. He holds or has applied for more than 250 US patents.

Dr. Yu is a Fellow of the ACM and a Fellow of the IEEE. He is associate editors of ACM Transactions on the Internet Technology and ACM Transactions on Knowledge Discovery in Data. He is a member of the IEEE Data Engineering steering committee and is also on the steering committee of IEEE Conference on Data Mining. He was the Editor-in-Chief of IEEE Transactions on Knowledge and Data Engineering (2001–2004), an editor, advisory board member and also a guest co-editor of the special issue on mining of databases. He had also served as an associate editor of Knowledge and Information Systems. In addition to serving as program committee member on various conferences, he will be serving as the general chairman of 2006 ACM Conference

on Information and Knowledge Management and the program chairman of the 2006 joint conferences of the 8th IEEE Conference on E-Commerce Technology (CEC' 06) and the 3rd IEEE Conference on Enterprise Computing, E-Commerce and E-Services (EEE' 06). He was the program chairman or co-chairs of the 11th IEEE International Conference on Data Engineering, the 6th Pacific Area Conference on Knowledge Discovery and Data Mining, the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, the 2nd IEEE International Workshop on Research Issues on Data Engineering: Transaction and Query Processing, the PAKDD Workshop on Knowledge Discovery from Advanced Databases, and the 2nd IEEE International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems. He served as the general chairman of the 14th IEEE International Conference on Data Engineering and the general co-chairman of the 2nd IEEE International Conference on Data Mining. He has received several IBM honors including 2 IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, 2 Research Division Awards and the 84th plateau of Invention Achievement Awards. He received an Outstanding Contributions Award from IEEE International Conference on Data Mining in 2003 and also an IEEE Region 1 Award for "promoting and perpetuating numerous new electrical engineering concepts" in 1999. Dr. Yu is an IBM Master Inventor.



Richard R. Muntz is a Professor and past chairman of the Computer Science Department, School of Engineering and Applied Science, UCLA. His current research interests are sensor rich environments, multimedia storage servers and database systems, distributed and parallel database systems, spatial and scientific database systems, data mining, and computer performance evaluation. He is the author of over one hundred and fifty research papers.

Dr. Muntz received the BEE from Pratt Institute in 1963, the MEE from New York University in 1966, and the Ph.D. in Electrical Engineering from Princeton University in 1969. He is a member of the Board of Directors for SIGMETRICS and past chairman of IFIP WG7.3 on performance evaluation. He was a member of the Corporate Technology Advisory Board at NCR/Teradata, a member of the Science Advisory Board of NASA's Center of Excellence in Space Data Information Systems, and a member of the Goddard Space Flight Center Visiting Committee on Information Technology. He recently chaired a National Research Council study on "The Intersection of Geospatial Information and IT" which was published in 2003. He was an associate editor for the Journal of the ACM from 1975 to 1980 and the Editor-in-Chief of ACM Computing Surveys from 1992 to 1995. He is a Fellow of the ACM and a Fellow of the IEEE.