# DEMON: Mining and Monitoring Evolving Data


Venkatesh Ganti* — *UW-Madison* — *vganti@cs.wisc.edu*

Johannes Gehrke† — *Cornell University* — *johannes@cs.cornell.edu*

Raghu Ramakrishnan‡ — *UW-Madison* — *raghu@cs.wisc.edu*



**Abstract**

Data mining algorithms have been the focus of much research recently. In practice, the input data to a data mining process resides in a large data warehouse whose data is kept up-to-date through periodic or occasional addition and deletion of blocks of data. Most data mining algorithms have either assumed that the input data is static, or have been designed for arbitrary insertions and deletions of data records.

In this paper, we consider a dynamic environment that evolves through systematic addition or deletion of *blocks* of data. We introduce a new dimension called the *data span dimension*, which allows user-defined selections of a temporal subset of the database. Taking this new degree of freedom into account, we describe efficient model maintenance algorithms for frequent itemsets and clusters. We then describe a generic algorithm that takes any traditional incremental model maintenance algorithm and transforms it into an algorithm that allows restrictions on the data span dimension. We also develop an algorithm for automatically discovering a specific class of interesting block selection sequences. In a detailed experimental study, we examine the validity and performance of our ideas on synthetic and real datasets.



**Keywords:** Data Mining, dynamic databases, evolving data, trends.


## 1 Introduction

Organizations have realized that the large amounts of data they accumulate in their daily business operations can yield useful "business intelligence," or strategic insights, based on observed patterns of activity. There is an increasing focus on *data mining*, which has been defined as the application of data analysis and discovery algorithms to large databases with the goal of discovering (predictive) models [FPSSU96]. Several


*Supported by a Microsoft Research Fellowship


†Work done when the author was at UW-Madison


‡This research was supported by Grant 2053 from the IBM corporation.

algorithms have been proposed for computing novel models, for more efficient model construction, to deal with new data types, and to quantify differences between datasets.

Most data mining algorithms so far have assumed that the input data is static and do not take into account that data evolves over time. Recently, the problem of mining evolving data has received some attention and incremental model maintenance algorithms for several data mining models have been developed [CHNW96, CVB96, FAAM97, TBAR97, EKS$^+$98, GGRL99b]. These algorithms are designed to incrementally maintain a data mining model under arbitrary insertions and deletions of records to the database.

But real-life data often does not evolve in an arbitrary way. Consider a data warehouse, a large collection of data from multiple sources consolidated into a common repository, to enable complex data analysis [CD97]. The data warehouse is updated with new batches of records at regular time intervals, e.g., every day at midnight. Thus the data in the data warehouse evolves through addition and deletion of batches of records at a time. We refer to data that changes through addition and deletion of "blocks" of records as *systematic (block) evolution*. A *block* is a set of records that are added simultaneously to the database. The main difference between arbitrary and systematic evolution is that in the former an individual record can be updated at any time, whereas in the latter blocks of records are added together. Also, all blocks in a systematically evolving database are logically ordered whereas in arbitrary evolution there is no order among tuples in a database.

In this paper, we assume a dynamic environment of systematically evolving data and introduce the problem of *mining systematically evolving data*. The main contributions of our work are:

1. We present a DEMONic[1] view of the world by exploring the problem space of mining systematically evolving data (Section 2). We introduce a new dimension called the *data span dimension*, which takes the temporal aspect of the data evolution into account and allows an analyst to "mine" relevant subsets of the data.

2. We describe new model maintenance algorithms with respect to the selection constraints on the data span dimension for two popular classes of data mining models: frequent itemsets and clustering (Section 3.1). These algorithms exploit the systematic block evolution to improve the state-of-the-art incremental algorithms. We also introduce a generic algorithm that takes any traditional incremental model maintenance algorithm and derives an incremental algorithm that allows restrictions on the data span dimension (Section 3.2). In particular, the generic algorithm can be instantiated with our incremental algorithms in Section 3.1.

[1]**D**ata **E**volution and **MON**itoring

3. We also address the problem of automatically discovering interesting selection constraints. Considering a class of constraints that identify sets of blocks with similar data characteristics, we propose an algorithm for discovering such constraints (Section 4).

4. In an extensive experimental study, we evaluate our algorithms on synthetic and real datasets, and compare them with previous work wherever possible (Section 5).

# 2 DEMON

In this section, we introduce the problem of mining systematically evolving data. We describe our model of systematic data evolution in Section 2.1. In Section 2.2, we enumerate the problem space of mining systematically evolving data by introducing the data span dimension, which allows temporal restrictions on the data being mined. Then we refine the type of restrictions by introducing the notion of a block selection sequence in Section 2.3.

## 2.1 Systematic Data Evolution

We now describe our model of evolving data. We use the term *tuple* generically to stand for the basic unit of information in the data, e.g., a customer transaction, a database record, or an $n$-dimensional point. The context usually disambiguates the type of information unit being referred to. A *block* is a set of tuples. We assume that the database $D$ consists of a (conceptually infinite) sequence of blocks $D_1, \ldots, D_k, \ldots$ where each block $D_k$ is associated with an *identifier* $k$. We assume without loss of generality that all identifiers are natural numbers and that they increase in the order of their arrival. Unless otherwise mentioned, we use $t$ to denote the identifier of the "latest" block $D_t$. We call the sequence of all blocks $D_1, \ldots, D_t$ currently in the database the *current database snapshot*.

Note that we do not assume that block evolution follows a regular period; different blocks may span different time units. For example, the first two blocks of data may be added to the database on Saturday and Sunday, respectively, and the third block on the following Friday. The framework can naturally handle this type of irregular block evolution. The lack of constraints on the time spanned by any block also allows us to incorporate hierarchies on the time dimension. (We just merge all blocks that fall under the same parent.)

## 2.2 Data Span Dimension

When mining systematically evolving data, some applications are interested in mining all the data accumulated thus far, whereas some other applications are interested in mining only a recently collected portion of the data.

As an example, consider an application that analyzes a large database of documents. Suppose the model extracted from the database through the data mining process is a set of document clusters, each consisting of a set of documents related to a common concept [Wil88]. The document cluster model is used to associate new, unclassified documents with existing concepts. Occasionally, a new block of documents is added to the database, necessitating an update of the document clusters. Typical applications in this domain are interested in clustering the entire collection of documents.

In a different application consider the database of the hypothetical `Demons'R Us` toy store which is updated daily. Suppose the set of frequent itemsets discovered from the database is used by an analyst to devise marketing strategies for new toys. The model obtained from all the data may not interest the analyst for the following reasons. (1) Popularity of most toys is short-lived. Part of the data is "too old" to represent the current customer patterns, and hence the information obtained from this part is stale and does not buy any competitive edge. (2) Mining for patterns over the entire database may dilute some patterns that may be visible if only the most recent window of data, say, the latest $28$ days, is analyzed. The marketing analyst may be interested in precisely these patterns to capitalize on the latest customer trends.

To capture these two different requirements, we introduce a new dimension, called the *data span dimension*, which offers two options. In the *unrestricted window (UW)* option, the relevant data consists of all the data collected so far. In the *most recent window (MRW)* option, a specified number $w$ of the most recently collected blocks of data is selected as input to the data mining activity. We call the parameter $w$ the *window size*; $w$ is application dependent and specified by the data analyst. Formally, let $D_1, \ldots, D_t$ be the current database snapshot. Then the *unrestricted window* (denoted $D[1,t]$) consists of all the blocks in the snapshot. If $t \geq w$ the *most recent window* (denoted $D[t-w+1,t]$) of size $w$ consists of the blocks $D_{t-w+1}, \ldots, D_t$; otherwise, it consists of the blocks $D_1, \ldots, D_t$. In the remainder of the paper, we assume without loss of generality that $t \geq w$. Our techniques can easily be extended for the special case $t < w$.

## 2.3 Block Selection Sequence

In this section, we introduce an additional selection constraint called the *block selection predicate* that can be applied in conjunction with the options on the data span dimension to achieve a fine-grained block selection.

The following hypothetical applications (of interest to a marketing analyst) defined on the `Demons'R Us` database motivate the finer-level block selection.

1. The analyst wants to model data collected on all Mondays to analyze sales immediately after the weekend. The required blocks are selected from the unrestricted window by a predicate that marks all blocks added to the database on Mondays.

2. The analyst is interested in modelling data collected on all Mondays in the past $28$ days (corresponding to the last $4$ weeks). In this case, a predicate that marks all blocks collected on Mondays in the most recent window of size $28$ selects the required blocks.

3. The analyst wants to model data collected on the same day of the week as today within the past $28$ days. The required blocks are selected from the most recent window of size $28$ by a predicate that, starting from the beginning of the window, marks all blocks added every seventh day.

Note that the block selection predicate is independent of the starting position of the window in the first and second applications whereas in the third application, it is defined relative to the beginning of the window and thus moves with the window. We now define the *block selection sequence (BSS)* to formalize the intuition behind the selection predicate. Informally, the BSS is a bit sequence of $0$'s and $1$'s; a $1$ in the position corresponding to a block indicates that the block is selected for mining, and a $0$ indicates that the block is left out.

**Definition 2.1** Let $D[1,t] = \{D_1, \ldots, D_t\}$ be the current database snapshot and let $D[t-w+1,t]$ be the most recent window of size $w$. A *window-independent block selection sequence* is a sequence $\langle b_1, \ldots, b_t, \ldots \rangle$ of $0/1$ bits. A *window-relative BSS* is a sequence $\langle b_1, \ldots, b_w \rangle$ of bits ($b_i \in \{0,1\}$), one per block in the most recent window. $\odot$

Note that the sharp distinction between the unrestricted window and the most-recent window allows the window-relative block selection sequence to exist. Otherwise, using a fixed block selection sequence and just the unrestricted window option, we cannot express the requirement of dynamically maintaining models on data "collected on all alternate days within the past 30 days."

**Automatic Block Selection Sequence Discovery:** Model maintenance with respect to a given block selection sequence assumes that the data analyst knows exactly what selection contraints need to be applied. However, in some cases the analyst may not be aware of such contraints. Even otherwise, the data analyst may want to know if the sequence of blocks contains any unknown interesting block selection sequences. We address this issue of automatically detecting interesting selection constraints in Section 4.

## 3 Incremental Model Maintenance Algorithms

In this section, we discuss incremental model maintenance algorithms for the two options on the data span dimension. In Section 3.1, we describe model maintenance algorithms for frequent itemsets and clustering under the unrestricted window option.[2] In Section 3.2, we describe a generic model maintenance algorithm called GEMM[3] for the most recent window option. The instantiation of GEMM requires a model maintenance algorithm for the unrestricted window option. The instantiated algorithm has identical performance characteristics (time between the arrival of a new block and the availability of the updated model) and main-memory requirements as the algorithm instantiating GEMM at the cost of a small amount of additional disk space and off-line processing. GEMM can be instantiated for any class of data mining models, and with any incremental model maintenance algorithm besides the ones we discuss in Section 3.1. Therefore, GEMM can take full advantage of specialized application-dependent incremental model maintenance algorithms to deliver better performance. Before describing our algorithms, we formally introduce the problems of frequent itemset computation and clustering. Since we do not describe any new algorithms for maintaining decision tree models, we do not discuss the decision tree models in detail.

**Set of Frequent Itemsets:** Let $\mathcal{I} = \{i_1, \ldots, i_n\}$ be a set of literals called *items*. A *transaction* and an *itemset* are subsets of $\mathcal{I}$. Each transaction is associated with a unique positive integer called the *transaction identifier*. A transaction $T$ is said to *contain* an itemset $X$ if $X \subseteq T$. Let $D$ be a set of transactions. The *support* $\sigma_D(X)$ of an itemset $X$ in $D$ is the fraction of the total number of transactions in $D$ that contain $X$: $\sigma_D(X) \stackrel{\text{def}}{=} \frac{|\{T : T \in D : X \subseteq T\}|}{|D|}$. Let $\kappa$ ($0 < \kappa < 1$) be a constant called the *minimum support*. An itemset $X$ is said to be *frequent* on $D$ if $\sigma_D(X) \geq \kappa$. The set of *frequent itemsets* $L(D, \kappa)$ consists of all itemsets that are frequent on $D$; formally, $L(D, \kappa) = \{X : X \subset \mathcal{I}, \sigma_D(X) \geq \kappa\}$. The *negative border* $NB^-(D, \kappa)$ of $D$ at minimum support threshold $\kappa$ is the set of all infrequent itemsets whose proper subsets are all frequent. Formally, $NB^-(D, \kappa) = \{X : X \subset \mathcal{I}, \sigma_D(X) < \kappa \wedge \forall Y \subset X, \sigma_D(Y) \geq \kappa\}$.

The *TID-list* $\theta_D(X)$ of an itemset $X$ is the list of transaction identifiers, sorted in the increasing order, of transactions in $D$ that contain the itemset $X$. The *size* of $\theta_D(X)$ is the (disk) space occupied by $\theta_D(X)$. We write $\theta(X)$ and $\sigma(X)$ instead of $\theta_D(X)$ and $\sigma_D(X)$ if $D$ is clear from the context.

[2]In prior work, we developed an algorithm for incremental decision tree construction [GGRL99b]. Hence we do not address this problem here.

[3]**GE**neric **M**odel **M**aintainer

**Clustering:** The clustering problem has been widely studied and several definitions for a cluster have been proposed to suit different target applications. In general, the goal of clustering is to find interesting groups (called *clusters*) in the dataset such that points in the same group are more similar to each other than to points in other groups. The notion of similarity between tuples is usually captured by a *distance* function, and the quality of clustering is usually measured by a distance-based *criterion function* (e.g., the weighted total or average distance between pairs of points in clusters). And, the goal of a clustering algorithm is to determine a good—as determined by the criterion function—partition of the dataset into clusters. A *cluster model* consists of all the clusters identified in the data. Since the clustering problem has been considered in several domains, many definitions exist which sometimes influence the algorithms as well. Without constraining ourselves to a specific approach, we adopt the following (semi-)formal definition from the Statistics literature for the clustering problem [JD88]. *Given the required number of clusters $K$, a dataset of $N$ points, a distance-based measurement function, and a criterion function, partition the dataset into $K$ groups such that the criterion function is optimized.*

## 3.1 Unrestricted Window

In this section, we describe incremental model maintenance algorithms for frequent itemsets and clusters for the unrestricted window option with respect to a user-specified BSS.

### 3.1.1 Set of Frequent Itemsets

When a new block $D_{t+1}$ is added to $D[1,t]$ and $b_{t+1} = 1$, the set of frequent itemsets needs to be updated. (If $b_{t+1} = 0$, the current set of frequent itemsets carries over to the new snapshot.) In this section, we discuss two new algorithms, called ECUT[4] and ECUT$^+$, for dynamically maintaining the set of frequent itemsets. These algorithms improve upon the previous best algorithm[5], BORDERS, which was independently developed by Feldman et al. [FAAM97] and Thomas et al. [TBAR97]. (The improvements exploit the systematic data evolution.) First, we briefly review the BORDERS algorithm before discussing new algorithms.

The BORDERS algorithm consists of two phases. (1) The *detection phase* recognizes that the set of frequent itemsets has changed. (2) The *update phase* counts a set of new itemsets required for dynamic maintenance. The detection phase relies on the maintenance of the negative border along with the set of frequent itemsets. When a new block $D_{t+1}$ is added to $D[1,t]$, the supports of the set of frequent itemsets $L(D[1,t],\kappa)$ and the negative border itemsets $NB^-(D[1,t],\kappa)$ are updated to reflect the addition. Detecting

[4]**E**fficient **C**ounting **U**sing **T**ID-lists

[5]Pudi et al. independently developed a new algorithm for maintaining frequent itemsets [PH00].

that a frequent itemset is no longer frequent is straightforward. The detection of new frequent itemsets is based on the following observation. If a new itemset $X$ becomes frequent on $D[1, t+1]$ then either $X$ or one of its subsets are in the negative border $NB^{-}(D[1,t],\kappa)$ of $D[1,t]$. Therefore, if there is no itemset $X \in NB^{-}(D[1,t],\kappa)$ whose support $\sigma(X)$ on $D[1,t+1]$ is greater than $\kappa$, then no new itemsets become frequent due to the addition of $D_{t+1}$, i.e., $L(D[1,t+1],\kappa) \subseteq L(D[1,t],\kappa)$.

The update phase is invoked if new frequent itemsets are flagged in the detection phase. Itemsets that are no longer frequent on $D[1,t+1]$ are deleted from $L(D[1,t],\kappa)$, new itemsets that are frequent on $D[1,t+1]$ are added to $L(D[1,t],\kappa)$. Deleting itemsets that are no longer frequent is straightforward.[6] If an itemset $X \in NB^{-}(D[1,t],\kappa)$ becomes frequent on $D[1,t+1]$, new candidate itemsets are generated by joining $X$ with $L(D[1,t],\kappa)$ (using the prefix join [AMS$^+$96]); after pruning those itemsets whose subsets are not frequent, the supports of the remaining candidate itemsets are counted. If a subset $L_X$ of the set of new candidates is frequent, then more new candidate itemsets are generated by joining $L_X$ with $L(D[1,t],\kappa) \cup L_X$, their support counted, and so on until no new frequent itemsets are found. The new set of frequent itemsets is added to the current set of frequent itemsets resulting in $L(D[1,t+1],\kappa)$. The set of new candidates (after the pruning step) is added to the negative border resulting in $NB^{-}(D[1,t+1],\kappa)$. Typically, the number of new candidate itemsets is very small [FAAM97, TBAR97]. The BORDERS algorithm counts the supports of new candidate itemsets by organizing them in a prefix tree [Mue95][7] and scanning the entire dataset $D[1,t]$. In this paper, we use the prefix tree data structure and refer to this counting procedure as *PT-Scan* (for **P**refix **T**ree-**Scan**).

## ECUT

To improve the support-counting algorithm in the update phase, we exploit systematic data evolution and the fact that, typically, only a very small number of new candidate itemsets needs to be counted. The intuition behind our new support-counting algorithm ECUT is similar to that of an index in that it retrieves only the "relevant" portion of the dataset to count the support of an itemset $X$. The relevant information consists of the set of TID-lists of items in $X$. ECUT uses TID-lists $\theta(i_1),\ldots,\theta(i_k)$ of all items in an itemset $X = \{i_1,\ldots,i_k\}$ to count the support of $X$. The cardinality of the result of the intersection of these TID-lists equals $\sigma(X)$. Since TID-lists, by definition, consist of transaction identifiers sorted in increasing order, the intersection can be performed easily; the procedure is exactly the same as the merge phase of a

[6] If an itemset $X \in L(D[1,t],\kappa)$ is no longer frequent on $D[1,t+1]$, then $X$ is deleted from $L(D[1,t],\kappa)$ and added to $NB^{-}(D[1,t],\kappa)$. All the supersets of $X$ are deleted from $NB^{-}(D[1,t],\kappa)$.

[7] A hash tree has also been proposed for the same purpose [AMS$^+$96].

sort-merge join. The support of an itemset $X = \{i_1, \ldots, i_k\}$ is given as follows:

$$\sigma_D(X) = \frac{|\{x : x \in \theta_D(i_1) \wedge x \in \theta_D(i_2) \wedge \ldots \wedge \theta_D(i_k)\}|}{|D|}$$

The size of the TID-list of an item $x \in \mathcal{I}$ is typically one to two orders of magnitude smaller than the size of $D$. The amount of data fetched by ECUT to count the support of an itemset $X = \{i_1, \ldots, i_k\}$ is equal to the sum of the supports $\sum_{j=1}^{k} \sigma(i_j)$ of all items in $X$ which, again, is typically an order of magnitude smaller than the space occupied by $D$. Therefore, whenever the number of itemsets to be counted is not large, ECUT is significantly faster than (previous) support-counting algorithms which scan the entire dataset $D[1, t]$.

**Organization of TID-lists:** To take full advantage of the TID-lists of items, we selectively read only the relevant portion of the TID-lists derived from the set of blocks selected by the BSS. The following two observations allow the TID-list of an item with respect to $D[1, t]$ to be partitioned into $t$ parts, one per block.

1. **Additivity property:** The support of an itemset $X$ on $D[1, t]$ is the *sum* of its supports in blocks $D_1, \ldots, D_t$.

2. $0/1$ **property:** Because of the nature of the block selection sequence that it either selects a block completely or not at all, we never need to count the support of an itemset $X$ partially in any block $D_i$, $i \in \{1, \ldots, t\}$.

The implication of the above two properties is that for $x \in \mathcal{I}$, all the TID-lists $\theta_{D_i}(x)$ for each block $D_i$ can be constructed when $D_i$ is added to the database and used—without any further changes—for counting supports of itemsets. Since the identifiers of transactions increase in the order of their arrival, materialization of TID-lists of items is straightforward. A block $D_i$ is scanned and the identifier of each transaction $T \in D_i$ is appended to $\theta_{D_i}(x)$ if $T$ contains the item $x$. The TID-lists of all items are materialized simultaneously by maintaining a buffer for each TID-list and flushed to disk whenever it is full.

**Space Requirements:** The space required to maintain TID-lists for all items in $\mathcal{I}$ is given by the sum of supports of all items in $\mathcal{I}$, which equals the space occupied by the database stored as a set of transactions. Moreover, any information that can be obtained from the transactional format can also be obtained from the set of TID-lists. Therefore, the TID-list representation is an alternative for the traditional transactional representation of the database; we no longer require the database in the traditional transactional format.

### ECUT$^{+}$

We now describe ECUT$^{+}$ which improves upon ECUT if additional disk space is available. The intuition behind ECUT$^{+}$ is that the support of an itemset $X$ can also be counted by joining the TID-lists of itemsets $\{X_1, \ldots, X_k\}$ as long as $X_1 \cup \cdots \cup X_k = X$, where the sizes of some or all $X_i$'s are greater than one. The greater the sizes of $X_i$'s, the faster it is to count the support of $X$ because the support of $X_i$ and hence the size of its TID-list typically decreases with increase in $|X_i|$; moreover, fewer TID-lists are sufficient to count the support of $X$. Therefore, if we materialize TID-lists of itemsets of size greater than one in addition to the TID-lists of single items, then the support of $X$ may be counted faster than using the TID-lists of the individual items in $X$. We now discuss the trade-offs involved as well as our solution.

For a block $D_i$, after materializing TID-lists of individual items, suppose an additional amount of disk space $M_i$ is available to materialize TID-lists of itemsets of size greater than one. How do we choose the appropriate set of itemsets whose TID-lists are to be materialized? Each TID-list $\theta_{D_i}(X)$ has a certain benefit and a cost. That $\theta_{D_i}(X)$ can be used to count the support of any itemset $Y \supset X$ adds to its benefit, and the cost of $\theta_{D_i}(X)$ is the space it occupies. However, to count the support of an itemset $Y$ we need a set of TID-lists of itemsets $Y_1, \ldots, Y_k$ such that $Y_1 \cup \cdots \cup Y_k = Y$, some of which could correspond to individual items. The goal now is to maximize the total benefit given an upper bound on the cost. This problem is the same as the NP-hard view materialization problem (encountered in data warehousing) on AND-OR graphs [Gup97]. Even the approximate greedy algorithm for selecting a set of itemsets that leads to a high benefit is, in the worst case, exponential in the number of materializable itemsets [Gup97]. Due to the very high complexity of even an approximate solution, we devise a simple heuristic which, as confirmed by our experiments, works well in practice.

The intuition behind our heuristic is based on the following observations. A significant reduction in the time required to count the support of an itemset results from the use of 2-itemsets instead of 1-itemsets. Also, the support $\sigma_D(X)$ of an itemset is indicative of its benefit because it is more likely that an itemset with higher support will be a subset of a larger number of itemsets whose supports need to be counted in future. These observations motivate the following heuristic choice of itemsets to be materialized.

Let $D = D[1,t]$ be the current window. For a new block $D_{t+1}$, we materialize the TID-lists of the set of all frequent 2-itemsets in $L(D[1,t], \kappa)$. If the sum of supports on $D_{t+1}$ of all frequent 2-itemsets is greater than $M_{t+1}$, we choose as many 2-itemsets as possible; an itemset $X$ with a higher overall support $\sigma_D(X)$ is chosen before another itemset $Y$ with a lower support $\sigma_D(Y)$. This simple heuristic provides a good trade-off between the reduction in time for counting the support of itemsets and the high complexity of more complicated algorithms.

As the database evolves, the data analyst may want to change the minimum support threshold from $\kappa$ to $\kappa'$. It is trivial to update the set of frequent itemsets when $\kappa' > \kappa$, because $L(D, \kappa') \subseteq L(D, \kappa)$. When $\kappa' < \kappa$, we can use the BORDERS algorithm augmented with ECUT or ECUT$^+$ in the update phase. The performance of the improved BORDERS algorithm depends on the number of new itemsets whose support is to be counted. We empirically study the trade-off between using PT-Scan and ECUT or ECUT$^+$ in Section 5.

#### 3.1.2 Clustering

We now describe our extensions to the BIRCH clustering algorithm [ZRL96] to derive an incremental clustering algorithm called BIRCH$^+$. We first briefly review BIRCH before describing our extensions.[8]

BIRCH works in two phases. In the first phase, the dataset is summarized into "sub-clusters." The second phase merges these sub-clusters into the required number of clusters using one of several traditional clustering algorithms. (See [DH73, K.90] for an overview of several algorithms.) The intuition behind the pre-clustering approach taken by these algorithms is explained by the following analogy. Suppose each tuple describes the location of a marble. Given a large number of marbles distributed on the floor, these algorithms replace dense regions of marbles with tennis balls where each tennis ball is a *sub-cluster* of a cluster of marbles. The number of tennis balls is a controllable parameter and the space required for representing a tennis ball is much smaller than that required for representing the collection of marbles. Therefore, it is possible to cluster these tennis balls using one's own favorite clustering algorithm, e.g., K-Means.

More formally, in the first *pre-clustering* phase, the dataset is scanned once to identify a small set of sub-clusters $\mathcal{C}$ which are represented very concisely using their *cluster features (CFs)*. The set $\mathcal{C}$ discovered in the first phase fits easily into main memory. The second phase further analyzes $\mathcal{C}$ and merges some sub-clusters that are close to each other to form the user-specified number of clusters. Therefore, as long as tuples that are to be placed in the same cluster are assigned to sub-clusters that are close to each other, the end result after the second phase is the same. This tolerance to slight errors in the assignment of tuples to sub-clusters makes BIRCH robust to changes in the input order. Since the second phase works on the in-memory set $\mathcal{C}$, it is very fast. Hence, the first phase dominates the overall resource requirements.

Our straight-forward extension, BIRCH$^+$, exploits the facts that BIRCH is not sensitive to the input order of the data and that the set of sub-clusters is maintainable incrementally. Let $D[1, t]$ be the current

[8] See [ZRL96] for the complete description.

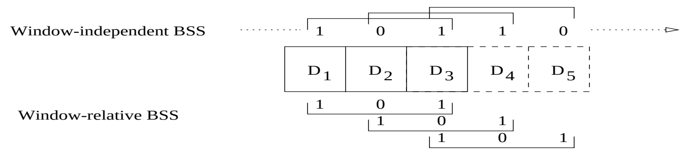


Figure 1: Most Recent Window

database snapshot. We give an inductive description of the algorithm. For the base case, $t = 1$, we just run BIRCH on $D[1, 1]$. At time $t + 1$, assume that the output of the first phase of BIRCH, the set of sub-clusters $\mathcal{C}_t$ is maintained in-memory. When $D_{t+1}$ is added to $D[1, t]$, we update $\mathcal{C}_t$ by scanning $D_{t+1}$ as if the first phase of BIRCH had been suspended and is now resumed. Let the updated set of sub-clusters be $\mathcal{C}_{t+1}$. We then invoke the second phase of BIRCH on $\mathcal{C}_{t+1}$ to obtain the user-specified number of clusters on $D[1, t + 1]$. The set $\mathcal{C}_{t+1}$ is maintained in-memory for the next block, completing the induction step.

Therefore, at any time $t$, the set of clusters is the same as if the non-incremental algorithm BIRCH was run on $D[1, t]$. Note that the response time of BIRCH$^+$ is very small, since the new block $D_{t+1}$ needs to be scanned only once and the second phase of BIRCH takes a negligible amount of time.

BIRCH$^+$ maintains the set of summarized cluster representations which, typically, is sufficient to discover and understand the sparse and dense regions in the dataset thus meeting the primary goal of clustering. However, if the set of points in the dataset $D[1, t + 1]$ needs to be partitioned based on their cluster membership, then we scan the dataset and associate with each point a label that corresponds to the cluster to which the point belongs. The second scan is characteristic of all clustering algorithms that use summarized cluster representations [ZRL96, GRS98, SCZ98].

### 3.2 Most Recent Window

We now describe GEMM, a generic model maintenance algorithm for the most recent window option. Given a class of models $\mathcal{M}$ and an incremental model maintenance algorithm $A_{\mathcal{M}}$ for the unrestricted window option, GEMM can be instantiated with $A_{\mathcal{M}}$ to derive a model maintenance algorithm (with respect to both window-independent and window-relative block selection sequences) for the most recent window option.

For both window-independent and window-relative block selection sequences, the central idea in GEMM is as follows. Starting with the block $D_{t-w+1}$, the window $D[t - w + 1, t]$ of size $w$ evolves in $w$ steps as each block $D_{t-w+i}$, $1 \leq i \leq w$, is added to the database. Therefore, the required model for the window $D[t - w + 1, t]$ can be incrementally evolved using $A_{\mathcal{M}}$ in $w$ steps. For example, the window $D[3, 5]$ in Figure 1 evolves in three steps starting with $D_3$, and consequently the model on $D[3, 5]$ can be built in three steps. The implication is that at any point, we have to maintain models for all future windows—windows which become current at a later instant $t' > t$—that overlap with the current window.

Suppose the current window $c_w$ is $D[t-w+1,t]$. There are $w-1$ future windows that overlap with $D[t-w+1,t]$. We incrementally evolve models (using $A_{\mathcal{M}}$) for all such future windows. For each future window $f_i = D[i+t-w+1, i+t]$, $0 < i < w$, we maintain the model with respect to an "appropriate" BSS for the prefix $D[i+t-w+1,t]$ of $f_i$ that overlaps with $c_w$. (The choice of the appropriate BSS for each prefix is explained later.) Since there are $w-1$ future windows overlapping with the current window, we maintain $w-1$ models in addition to the required model on the current window. Whenever a new block is added to the database shifting the window to $D[t-w+2,t+1]$, the model corresponding to the suffix $D[t-w+2,t]$ of $c_w$ is updated "appropriately" using $\mathcal{A}_{\mathcal{M}}$ to derive the required model on the new window $D[t-w+2,t+1]$.

As an example, consider the current database snapshot $D[1,3]$ with $w = 3$ in Figure 1. The future windows that overlap with $D[1,3]$ are $D[2,4]$ and $D[3,5]$. The models that are maintained in addition to the current model on $D[1,3]$ are extracted from $D[2,3]$ and $D[3,3]$ —the prefixes of $D[2,4]$ and $D[3,5]$ that overlap $D[1,3]$.

The choice of the BSS for extracting a model from the overlap between the current window and a future window depends on the type of BSS: window-independent or window-relative. We first describe the choice for the window-independent BSS and then extend it to the window-relative BSS.

### 3.2.1 Window-independent BSS

Consider the database snapshot $D[1,3]$ shown in Figure 1 with $w = 3$ and the window-independent BSS $\langle b_1, b_2, \ldots\rangle = \langle 10110\cdots\rangle$ (shown above the window). The current model on $D[1,3]$ is extracted from the blocks $D_1$ and $D_3$. After $D_4$ is added, the window shifts right and the new model on $D[2,4]$ is extracted from the blocks $D_3$ and $D_4$. We observe that the new model can be obtained by updating (using $A_{\mathcal{M}}$) the model extracted from $D_2$ and $D_3$ (the prefix of $D[2,4]$ that overlaps with $D[1,3]$). The observation here is that the relevant set of blocks (for the model extracted from $D[2,3]$) is selected from $D[1,3]$ by projecting the two bits $b_2$ and $b_3$ from the original BSS $\langle 10110\cdots\rangle$, and by padding the projection $b_2, b_3$ with a zero bit in the leftmost place to derive $\langle 0, b_2, b_3\rangle$. We call the operation of deriving a new BSS by projecting the relevant part from the window-independent BSS the *projection operation*.

We now formalize the projection operation. Without loss of generality, we use $D[1,w]$ (set $t = w$ in $D[t-w+1,t]$) to represent the current window of size $w$. Let $b = \langle b_1, \ldots, b_w, \ldots\rangle$ be the window-independent BSS. The *projection* operation takes as input a window-independent BSS $b$, the latest block identifier $t$, and a positive integer $k < w$ to derive a new sequence of length $w$ (the window size) that selects the relevant set of blocks (w.r.t. $b$) from the current window $D[1, w]$. Informally, the new sequence is the projection $b_{k+1}, \ldots, b_w$ from $b$ padded with $k$ zeroes in the $k$ leftmost places: $0, \ldots, 0, b_{k+1}, \ldots, b_w$. Formally, the *$k$-projected sequence* (denoted $b_k^w$) is given by $\langle b'_1, \ldots, b'_w \rangle$ where

$$b'_i \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if } 0 \le i \le k \\ b_i, & \text{if } k < i \le w \end{cases}$$

We need to introduce some more notation for describing the model maintenance. Let $m(D[1, w], b) \in \mathcal{M}$ denote the model extracted from the window $D[1, w]$ with respect to the BSS $b$. Let $A_{\mathcal{M}}(m, D_j)$ denote the updated model returned by $A_{\mathcal{M}}$ when a block of data $D_j$ is added to the dataset from which the model $m$ was extracted. Let $A_{\mathcal{M}}(D, \phi)$ represent the model extracted from the dataset $D$.

GEMM maintains a collection of models and updates it whenever a new block is added to the database. We now define the collection of models and describe the update operation.

**Collection of Models:** Given the current window $D[1, w]$ and the BSS $b = \langle b_1, \ldots, b_w, \ldots \rangle$, we maintain the collection $\mathcal{M}_b^{D[1,w]}$ of models defined as follows.

$$\mathcal{M}_b^{D[1,w]} = \{m(D[1, w], b_k^w) : k = 0, \ldots, (w - 1)\}$$

Informally, the collection consists (in addition to the currently required model) of a model for every future window overlapping with $D[1, w]$, and $b_k^w$ defines the BSS with respect to which the model is extracted from $D[1, w]$. Note that $m(D[1, w], b_0^w)$ is the required model on the current window $D[1, w]$ with respect to the BSS $b$.

---

**Algorithm 3.1 GAMMA-Update($A_{\mathcal{M}}$, $\mathcal{M}_b^{D[1,w]}$, $D[1, w]$, $b$, $D_{w+1}$)**

*/* Output:* $\mathcal{M}_b^{D[2,w+1]}$ */

**begin**

*Set* $\mathcal{M}_b^{D[2,w+1]} = \mathcal{M}_b^{D[1,w]} - \{m(D[1, w], b_0^w)\} \cup \{m(D_{w+1}, b_{w+1})\}$

**foreach** $k$ **in** $\{1 \ldots (w - 1)\}$

$$m(D[2, w + 1], b_{k-1}^{w+1}) = \begin{cases} A_{\mathcal{M}}(D_{w+1}, m(D[1, w + 1], b_k^w)) & \textit{if } b_{w+1} = 1 \\ m(D[1, w + 1], b_k^w) & \textit{if } b_{w+1} = 0 \end{cases}$$

**end** /* *foreach* */

**end**

---

**Updating the Collection of Models:** When a new block $D_{w+1}$ is added to the database the (most recent) window shifts to $D[2, w+1]$. Recall that each model in $\mathcal{M}_b^{D[1,w]}$ is extracted (with respect to an appropriate BSS) from the prefix of a future window. The addition of a new block extends these prefixes by one more block, and the models are updated to reflect this extension. The update operation on the collection of models $\mathcal{M}_b^{D[1,w]}$ is described in Algorithm 3.1.

The model $m(D[2, w+1], \langle b_2, \ldots, b_{w+1} \rangle)$ is the new model required with respect to the BSS $b = \langle b_1, \ldots, b_w, \ldots \rangle$ on the new window $D[2, w+1]$. For the example in Figure 1, $w = 3$ and the window-independent BSS is $\langle 10110 \rangle$. Therefore, the collection of models maintained for the window $D[1, 3]$ is:

$$\{m(D[1,3], \langle 101 \rangle), m(D[1,3], \langle 001 \rangle), m(D[1,3], \langle 001 \rangle)\}$$

When the new block $D_4$ is added, the collection of models is updated to:

$$\{m(D[2,4], \langle 011 \rangle) = A_{\mathcal{M}}(D_4, m(D[1,3], \langle 001 \rangle)), m(D[2,4], \langle 011 \rangle), m(D[2,4], \langle 001 \rangle)$$

Note that some of the models simultaneously maintained might be identical. For example, if $b_i^w = b_j^w$, then the models $m(D[1, w], b_i^w)$ and $m(D[1, w], b_j^w)$ are identical. In the above example, the second and third models in the collection of models on $D[1, 3]$ are identical. (Both are equal to $m(D[1, 3], \langle 001 \rangle)$.) Therefore, the actual number of different models maintained at any given time may be less than $w$.

### 3.2.2 Window-relative BSS

Consider the database snapshot $D[1, 3]$ shown in Figure 1 with $w = 3$ and the window-relative BSS=$\langle 101 \rangle$. The current model on $D[1, 3]$ is extracted from the blocks $D_1$ and $D_3$. When $D_4$ is added, the window shifts right and the new model on $D[2, 4]$ is extracted from the blocks $D_2$ and $D_4$. Observe that the new model can be obtained by updating (using $A_{\mathcal{M}}$) the model extracted from the block $D_2$. The important observation is that the relevant set of blocks (for extracting the model from the overlap between $D[1, 3]$ and $D[2, 4]$) is selected from $D[1, 3]$ by the BSS $\langle 010 \rangle$—obtained by right-shifting the original BSS $\langle 101 \rangle$ once and padding the leftmost bit with a zero. We call this operation the *right-shift* operation.

The *right-shift* operation takes as input a window-relative BSS $b$, the current time stamp, and a positive integer $k$ ($k < w$) to derive a new sequence of length $w$ that selects the relevant set of blocks (w.r.t. $b$). Informally, the relevant set of blocks correspond to the set chosen by sliding $b$ forward by $k$ blocks, padding the leftmost $k$ bits with zeroes, and truncating the sequence that slides beyond $D_w$. Formally, if $b = \langle b_1, \ldots, b_w \rangle$ then the $k$-right-shifted sequence is $\langle b'_1, \ldots, b'_w \rangle$ where

$$b'_i \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if } 0 \leq i \leq k \\ b_{(i-k)}, & \text{if } k < i \leq w \end{cases}$$

The procedure for maintaining and updating a collection of models for a window-relative BSS is analogous to Algorithm 3.1 with the $k$-right-shift operation substituted for the $k$-project operation.

### 3.2.3 Response Time and Space Requirements

In this section, we denote the model on the window $D[1, w]$ with respect to a (window-independent or window-relative) BSS $b$ by $m(D[1, w], b)$. We define the *response time* to be the time elapsed between the addition of a new block $D_{w+1}$ and the availability of the updated model $m(D[2, w+1], b)$. From Algorithm 3.1, we observe that for either type of BSS, the computation of the new model $m(D[2, w+1], b)$ involves at most a single invocation of $\mathcal{A}_{\mathcal{M}}$ with the two arguments: $D_{w+1}$ and $m(D[2, w], b')$ (where $b'$ is defined by the projection or the right-shift operations). Therefore, the response time is less than or equal to the time taken by $A_{\mathcal{M}}$ to update the model $m(D[2, w], b')$ with $D_{w+1}$.

Except for the model $m(D[2, w+1], b)$, the models in $\mathcal{M}_b^{D[2,w+1]}$ are not required immediately in the new window. Therefore, these updates are not time-critical and can be performed off-line when the system is idle. However, some of these models need to be updated before the subsequent block arrives. An important implication of the lack of immediacy of these updates is that the collection $\mathcal{M}_b^{D[1,w]}$ of models except $m(D[2, w], b')$ can be stored on disk and retrieved when necessary. Thus main memory is not a limitation as long as a single model fits in-memory. Like all current data mining algorithms, we assume that at least one model fits into main memory. In general, we maintain $w - 1$ additional models on disk. Since the space occupied by a model is insignificant when compared to that occupied by the data in each block, the additional disk space required for these models is negligible.

### 3.2.4 Options and Optimizations

Certain classes of models are also maintainable under deletion of tuples. For example, the frequent itemsets model can be maintained under deletions of transactions. The algorithm proceeds exactly as for the addition of transactions except that the support of all itemsets contained in a deleted transaction is decremented. Maintainability under deletions gives two choices for model maintenance under the most recent window option. (1) GEMM instantiated with the model maintenance algorithm $A_{\mathcal{M}}$ for the addition of new blocks.

(2) $A^u_{\mathcal{M}}$ that directly updates the model to reflect the addition of the new block and the deletion of the oldest block in the current window. We first discuss the space-time trade-offs between the two choices for the special case when the BSS=$\langle 11\ldots 1\rangle$, and then for an arbitrary BSS.

Let the BSS be $\langle 1\ldots 1\rangle$. The first option GEMM requires slightly more disk space to maintain $w-1$ models. The response time is that of invoking $A_{\mathcal{M}}$ to add the new block. In the second option $A^u_{\mathcal{M}}$, we only maintain one model. However, $A^u_{\mathcal{M}}$ has to reflect the addition of the new block and the deletion of the oldest block and hence approximately (assuming that deletion of a tuple takes as much time as addition and the blocks being deleted and added are of the same size) takes twice as long as GEMM. Therefore, GEMM has better response time characteristics with a small increase in disk space requirements.

The full generality of GEMM comes to the fore for classes of models that cannot be maintained under deletions of tuples, and in cases where model maintenance under deletion of tuples is more expensive than that under insertion. For instance, the set of sub-clusters in BIRCH cannot be maintained under deletions, and the cost incurred by incremental DBScan to maintain the set of clusters when a tuple is deleted is higher than that when a tuple is inserted [EKS$^+$98].

When we consider an arbitrary BSS, a major drawback of using $A^u_{\mathcal{M}}$ to maintain models on the most recent window with respect to an arbitrary window-relative BSS is that it may require deletion and addition of many blocks to update the model. Recall that a (window-relative) BSS chooses a subset $B$ of the set of blocks $\{D_1,\ldots,D_w\}$ in the window. When the window shifts right, depending on the BSS, a number ($\geq 1$) of blocks may be newly added to $B$ and more than one block may be deleted from $B$. Therefore, $A^u_{\mathcal{M}}$ scans all blocks in the newly added set as well as the deleted set. For certain block selection sequences, it may reduce to the naive reconstruction of the model from scratch as illustrated by the following example. Let the current database snapshot be $D[1,10]$, and the window-relative BSS be $\langle 1010101010\rangle$. The current model is constructed from $\{D_1,D_3,D_5,D_7,D_9\}$. If the window shifts right then the new set of blocks is $\{D_2,D_4,D_6,D_8,D_{10}\}$, which is disjoint from the earlier set.

# 4 Pattern Detection

In the previous section, we discussed model maintenance algorithms for a dynamically varying subset of the database as specified by an arbitrary block selection sequence. The evolutionary nature of the data also opens up new problems. We can ask how the behavior, over time, of data characteristics changes. For example, do data exhibit cyclic or seasonal patterns? Note that this problem of *pattern detection* is not tied to our notion of systematic data evolution, but arises for any dynamically changing database. However, we can always view such a data repository as a (logical) sequence of blocks. Therefore, assuming our model of systematic block evolution, we now describe some results for detecting "patterns of similar blocks." In the language of block selection sequences, our approach, intuitively, is to identify a set of block selection sequences where all blocks of data within each BSS are similar in their data characteristics.

As a motivating example, consider the brand manager of a brand of frozen pizza. At any time, she needs accurate predictions of sales for the upcoming weeks in order to coordinate production, distribution and marketing. Specifically, the manager would like to use historic sales information to discover base line sales trends that can be used to predict sales in upcoming weeks. Simple patterns like "the number of pizzas sold in the two weeks before Superbowl Sunday is significantly higher" are probably known to the manager and knowledge of such a folklore pattern will not result in any competitive advantage. However, lack of knowledge of such patterns, or ignorance of common patterns, will be a striking competitive disadvantage. Model maintenance with respect to a block selection sequence addressed this problem of maintaining models for known interesting patterns. A more exploratory question is: how can we discover interesting patterns (or, equivalently, block selection sequences) of similar blocks in systematically changing data?

To detect a pattern from a sequence of blocks, we require a notion of similarity between any two blocks of data. In prior work [GGRL99a], we developed the FOCUS framework for computing an interpretable, statistically qualifiable measure of difference called *deviation* between two datasets. The deviation quantifies the difference between interesting characteristics in each dataset as reflected in the data mining models they induce. The deviation framework can be instantiated with any one of three popular data mining models: frequent itemsets, decision tree classifiers, and clusters. The central idea is that a broad class of models can be described in terms of a *structural component* and a *measure component*. The structural component identifies "interesting regions," and the measure component summarizes the subset of the data that is mapped to each region. Given two datasets and models induced from these datasets, the framework extends the structural components of the two models to a common structural component to reconcile the differences between them. Now, the deviation between them is computed as the aggregate between the measures of the two datasets over all regions in the common structural component. The computation of the deviation measure is fast since it requires at most one scan of each dataset. (See [GGRL99a] for details.)

In the remainder of this section, we consider a fixed class of data mining models $\mathcal{M}$ and denote by $\delta_{\mathcal{M}}(D_1, D_2)$ the deviation value between two datasets $D_1$ and $D_2$ through the class of models $\mathcal{M}$. The measure of similarity between two datasets $D_1$ and $D_2$ is the statistical significance of the deviation $\delta_{\mathcal{M}}(D_1, D_2)$ between $D_1$ and $D_2$. Informally, the statistical significance of the deviation is the probability that both datasets are drawn from the same underlying hypothetical process generating data. Formally, we say that blocks $D_1$ and $D_2$ are *$\mathcal{M}$-similar at significance level* $\alpha$ ($0 < \alpha < 1$) if $\delta_{\mathcal{M}}(D_1, D_2) < \alpha$. In practice, this similarity function is used with a binary range: $0$ or $1$ where the function takes a value $1$ if the two blocks are similar and $0$, otherwise. Note that our notion of similarity is symmetric, but not transitive.

Given that we have a similarity function between any pair of blocks, one approach for finding groups of similar blocks is to treat each block as an object and then discover clusters of (similar) objects. (Several clustering algorithms may be applicable here.) This approach has the following drawback. Most clustering algorithms partition the set of objects and do not allow overlap between clusters [DH73, K.90]. Thus individual sequences of blocks corresponding to different clusters do not overlap, which is a very strong restriction. For instance, the two patterns "blocks collected every Monday," and "blocks collected on the first day of every month" may not co-exist. For this reason, the clustering formulation is not suitable for the problem of identifying patterns of similar blocks. We now discuss a simple alternate formulation that overcomes this problem, and allows an efficient algorithm (unlike the NP-hard clustering problem).

To allow individual patterns represented by block selection sequences to overlap and to explicitly take the logical order among blocks into account, we introduce the notion of *compact sequences*. A compact sequence is a maximal sequence of pairwise similar blocks such that any block between the first and last blocks in the sequence that is similar to each block before it in the sequence also belongs to the sequence. We call such sequences compact because they do not leave out any block that is eligible to be included in the sequence. In other words, there are no "holes" in these sequences. We realize that the set of compact sequences may not include all classes of interesting sequences. However, we believe that the set of compact sequences may be analyzed further to discover specialized types of patterns by placing additional constraints like *cyclicity* on the set of blocks in a compact sequence. Such specialized types of sequences can be computed by subjecting the set of sequences to a post-processing step. For instance, if $\langle D_1, D_3, D_4, D_5, D_7 \rangle$ is a compact sequence, we can easily derive the cyclic sequence $\langle D_1, D_3, D_5, D_7 \rangle$ from this input sequence.

For the sake of clarity in presentation, we use the following notation. Let $\mathcal{M}$ be a class of models—selected from among frequent itemset models, decision tree models, and cluster models—to instantiate the deviation framework. (For instance, the analyst may select a particular class.)

**Definition 4.1** Let $f$ and $g$ be the difference and aggregation functions in the instantiation of a deviation function. Let $\alpha_{\mathcal{M}}(D_1, D_2)$ denote the statistical significance of the deviation between two blocks $D_1$ and $D_2$. We say that blocks $D_1$ and $D_2$ are *$\mathcal{M}$-similar at significance level* $\alpha$ ($0 < \alpha < 1$) if $\delta_{\mathcal{M}}(D_1, D_2) < \alpha$.

We call a sequence $S$ of blocks $\{D_{i_1}, \ldots, D_{i_k}\}$ *compact* if (1) each pair of blocks in $S$ are similar, and

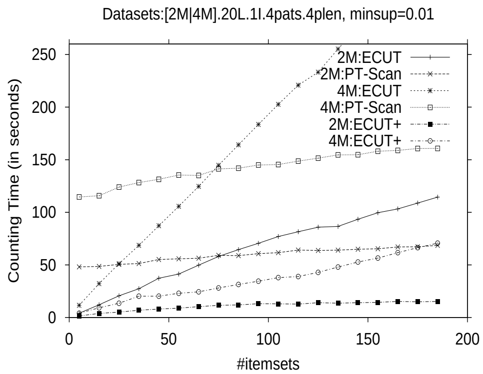


Figure 2: Counting Times

| Datasets | $\kappa$ | % extra space for freq. 2-itemsets |
|---|---|---|
| {2M, 4M}.20L.1I.4pats.4plen | 0.008 | 25.3 |
| {2M, 4M}.20L.1I.4pats.4plen | 0.010 | 11.8 |
| {2M, 4M}.20L.1I.4pats.4plen | 0.012 | 5.3 |

Figure 3: % Extra Space for ECUT$^+$

(2) for any block $D_i \notin S$, with an identifier $i$ between $i_1$ and $i_k$, $D_i$ is not similar to at least one block $D_j$ in $S$ where $i_1 \leq j < i$. $\odot$

Consider the sequence of blocks $\{D_1, D_2, D_3, D_4\}$ where only the pairs $(D_1, D_2)$, $(D_1, D_3)$, $(D_1, D_4)$, and $(D_2, D_4)$ are similar. Then the sequence $\{D_1, D_2, D_4\}$ is compact whereas the sequences $\{D_1, D_2, D_3\}$ (which violates (1)) and $D_1, D_4$ (which violates (2)) are not.

We now describe a simple algorithm that incrementally computes all compact sequences of blocks when the unrestricted window option on the data span dimension is selected.[9] The basic idea is to incrementally maintain the set of compact sequences as new blocks are added to the database. We give an inductive description of the algorithm. In the base case ($t = 1$), $D_1$ is added to the (empty) database. The set of compact sequences just consists of the single sequence $D_1$. For the induction step ($t > 1$), assume that there are exactly $t$ compact sequences $\mathcal{G}_t = G_1, \ldots, G_t$ in $D[1, t]$. Let $D_{t+1}$ be the block that is added at time $t + 1$. We set $G_{t+1} = \{D_{t+1}\}$ and we extend each $G_i$ with $D_{t+1}$ if the extended sequence is still compact. We set $\mathcal{G}_{t+1} = G_1, \ldots, G_{t+1}$, completing our description of the algorithm.

To avoid repeated computation of deviations between the same pair of blocks, we maintain a matrix consisting of all pair-wise deviations in the current database snapshot $D[1, t]$. Whenever a new block $D_{t+1}$ is added, we augment the matrix with the values of $\delta_{\mathcal{M}}(D_{t+1}, D_i)$ for $i \in \{1, \ldots, t\}$. It is straighforward to show that the above algorithm actually computes all compact sequences.

[9]This algorithm can be extended easily to apply to the most recent window option.

# 5 Performance Evaluation

In this section, we first evaluate the performance of our incremental model maintenance algorithms for the unrestricted window option. Since the response time for model maintenance under the most recent window option using GEMM is the same as the response time for model maintenance under the unrestricted window option, experimental results for the most recent window option are subsumed by results from the unrestricted window option. We then show results from applying our definition of compact block selection sequences and the corresponding algorithm to a real dataset of web proxy traces. All running times were measured on a 200 MHz Pentium Pro PC with 128 MB of main memory, and running solaris 2.6.

## 5.1 ECUT and ECUT$^+$

In this set of experiments, we compared the running time of ECUT and ECUT$^+$ with the running time of BORDERS [FAAM97, TBAR97]. Incremental maintenance of large itemsets proceeds in two phases, and the detection phase of our algorithms is identical to the detection phase of BORDERS. Thus we first measured the performance improvements of our techniques restricted to the update phase, and then examine how much each phase contributes to the overall model maintenance time.

We used the data generator developed by Agrawal et al. [AS94] to generate synthetic data. We write $N$M.$t_l$L.$|\mathcal{I}|$I.$N_p$pats.$p$plen to denote a dataset with $N$ million transactions, an average transaction length $t_l$, $|\mathcal{I}|$ items (in multiples of $1000$'s), $N_p$ patterns (in multiples of $1000$'s), and average pattern length $p$. The running times of Algorithm ECUT$^+$ had all 2-frequent itemsets in each block materialized, thus facilitating the best performance improvements. We observed in our experiments, that for the ranges of the minimum support thresholds and dataset parameters that we considered, the additional amount of space required for this materialization was less than $25\%$ of the overall dataset size (see Figure 3).

**Experiment 1:** We compared the update phases of ECUT and ECUT$^+$ with the update phase of BORDERS, called PT-Scan. We computed a set of frequent itemsets at the $1\%$ minimum support from the dataset $\{2, 4\}$M.20L.1I.4pats.4plen, then randomly selected a set of itemsets $S$ from the negative border and counted the support of all itemsets $X \in S$ against $D$. We varied the size of $S$ from $5$ to $180$. Figure 2 shows that all algorithms scale linearly with the number of itemsets in $S$, and the size of the input dataset $D$. ECUT outperforms PT-Scan when $|S| < 75$, and ECUT$^+$ outperforms PT-Scan in the entire range considered. When $|S| < 40$, ECUT is more than twice as fast as PT-Scan and ECUT$^+$ is around $8$ times as fast as PT-Scan. (Our results and previous work [FAAM97, TBAR97] show that $|S|$ is typically less than $30$. We

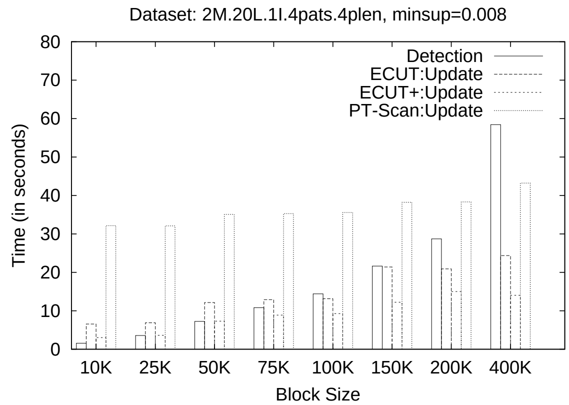


Figure 4: $*$M.20L.1I.8pats.4npl, $\kappa = 0.008$

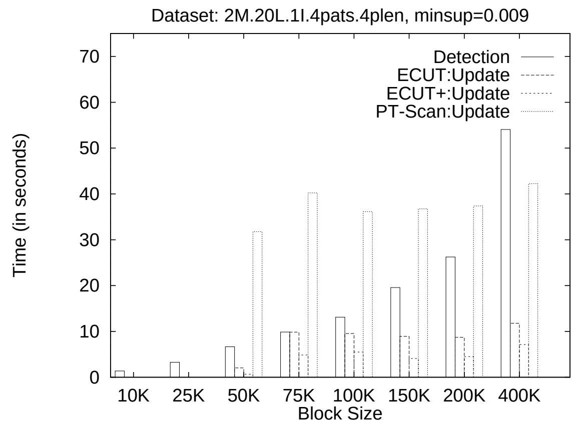


Figure 5: $*$M.20L.1I.8pats.4npl, $\kappa = 0.009$

considered large $|S|$ to thoroughly explore the tradeoffs between the algorithms.)

**Experiment 2:** We compared the total time taken by the algorithms, broken down into detection phase and update phase. We first computed the set of frequent itemsets at a certain minimum support threshold $\kappa$ from a first block. We then measured the overall maintenance time required to update the frequent itemsets when a second block is added. We fixed the distribution parameters for the first block to be 2M.20L.1I.4pats.4plen, and varied the value of $\kappa$ and the distribution parameters for the second block as follows. $\kappa$ is chosen from two values: $0.008$ and $0.009$. The second block is generated with parameter settings $*$M.20L.1I.8pats.4plen (first set) and $*$M.20L.1I.4pats.5plen (second set). The distribution characteristics in the second set of parameters causes more changes in the set of frequent itemsets. Besides these distribution parameters, we also varied the number of transactions in the second block from 10K to 400K ($0.5\% - 20\%$ of the first block's size).

The results from the first set of parameters are shown in Figures 4 and 5, and the results from the second set in Figures 6 and 7. First, note that the update phase of BORDERS dominates the overall maintenance time. Second, in most cases, ECUT and ECUT$^+$ are significantly faster than PT-Scan. When the sizes of the new (second) block are reasonably small relative to the old (first) block (less than $5\%$ of the original dataset size), our algorithms are between $2$ to $10$ times faster than PT-Scan, reducing the maintenance cost sometime by an order of magnitude. In general, whenever ECUT or ECUT$^+$ were used in the update phase, the detection phase dominates the total maintenance time, whereas for BORDERS the reverse is true.

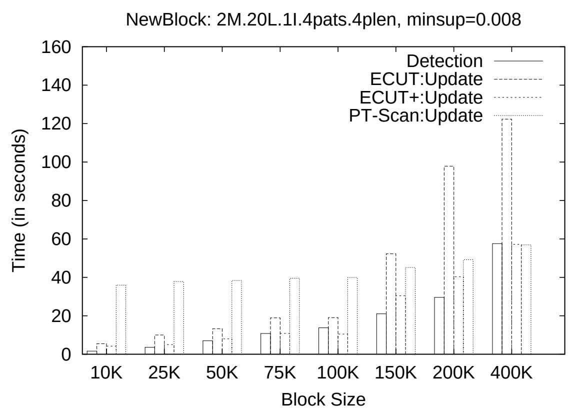


Figure 6: $*$M.20L.1I.4pats.5npl, $\kappa = 0.008$

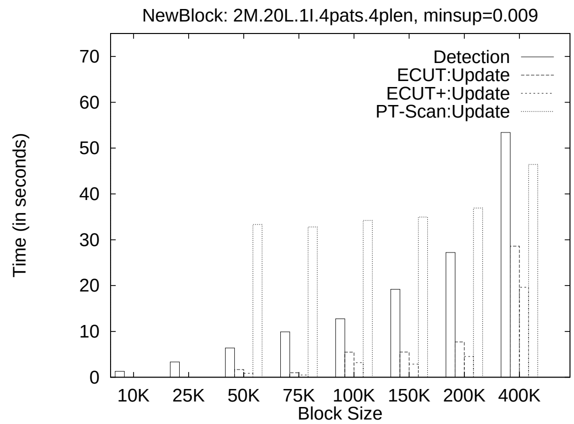


Figure 7: $*$M.20L.1I.4pats.5npl, $\kappa = 0.009$

## 5.2 BIRCH$^+$

In this section, we compare the running times of BIRCH$^+$ and the non-incremental standard version of BIRCH, which clusters the entire database whenever a new block arrives. Since Zhang et al. [ZRL96] showed that the output of BIRCH is not sensitive to the input order, we do not present any results on order-independence. For the experiments in this section, we used the synthetic data generator described by Agrawal et al. [AGGR98]. We generated clusters distributed over all dimensions.[10] The synthetic data generator requires three parameters: the number of points $N$ in multiples of millions, the number of clusters $K$, and the dimensionality $d$. A dataset generated with this set of parameters is denoted $N$M.$K$c.**d**d.

We present an experiment on a representative dataset chosen from a wide variety of datasets we experimented with. The results are similar for all datasets. We consider two blocks of data: $1$M.$50$c.$5$d and $*$M.$50$c.$5$d. We varied the number of tuples in the second block between $100$K and $800$K and added $2\%$ uniformly distributed noise points to perturb the cluster centers. Figure 8 shows that BIRCH$^+$ significantly outperforms BIRCH.

## 5.3 Pattern Detection on Web Proxy Traces

Through our experiments in this section, we test the validity of our notion of a compact sequence of blocks by examining results on a real dataset. Detection of interesting sequences would be a validation that the restriction of compactness is not unnatural.

Our real dataset is a set of web proxy traces collected at DEC [TK]. It consists of more than $22$ million

[10]In general, the data generator can also generate clusters over a subset of the set of all dimensions.

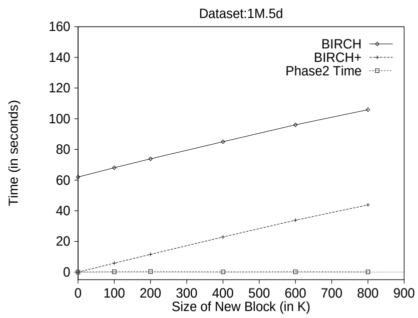


Figure 8: BIRCH$^+$

| Granularity | Trend |
|---|---|
| 24 hr | All working days except $9-9-1996$ |
| 12 hr | 12 Noon - 12 PM on all working days |
| 8 hr | 8 AM - 4 PM on all working days except $9-9-1996$ |
| 8 hr | 4 PM - 12 PM on all Tuesdays and Thursdays |
| 6 hr | 12 Noon - 6 PM on all working days except $9-9-96$ |
| 4 hr | 12 Noon - 4 PM on all working days except $9-9-96$ |
| 4 hr | 4 PM - 8 PM on all Tuesdays and Thursdays |

Figure 9: Patterns discovered in the Web Proxy Traces

tuples of web page requests collected over a period of 21 days between 8 AM on $9-2-1996$ and 12 PM on $9-22-1996$. Besides other information, the tuple of each web page request consists of the following fields: a timestamp, the type of the object requested (e.g., gif, jpg, etc.), and the number of bytes in the response. The requested objects are classified into 10 different types and we discretized the number of bytes received into 1000 consecutive intervals of size 10000 bytes. Our goal was to model potential relationships between the types of request and the number of bytes received. Thus we treated each tuple as a transaction consisting of the object type and the bucket number of the response size, and chose as data mining model the set of frequent itemsets at a minimum support level of 1%. Using the timestamp field in the database, we segmented the dataset into blocks while varying the block size at five different granularities (4, 6, 8, 12, and 24 hour intervals).

Figure 9 summarizes some of the patterns discovered in the database. From an analyst's point of view, there are reasonable explanations for each pattern shown. Besides the patterns shown in Figure 9, we also find other "surprising" information. For instance, the data traced on Monday, 9-9-1996, is significantly different from the data traced on other working days. The statistical significance of the deviation values are as high as 99% and our pattern detection algorithm recognizes this unusual block and does not include it in any of the currently maintained patterns (some of which are shown in Figure 9). The patterns shown exclude the weekends and the holiday $9-2-1996$ (labor day), thus recognizing the dissimilarity between data collected on weekdays and weekends. The following sequence (obtained for a block granularity of 4 hours) illustrates that late night weekday blocks can be similar to blocks on weekends: $\langle$[8PM-12PM] on 9-5-1996 and [0AM-4AM] on 9-6-1996 (Thursday to Friday night), [12Noon-4PM] on 9-7-1996 (Saturday afternoon), [8PM-12PM] on 9-18-1996 (Wednesday night), [4AM-8AM] on 9-20-1996 (very early Friday morning) $\rangle$.

From these qualitative results with the real dataset, we draw the following two conclusions. First, the

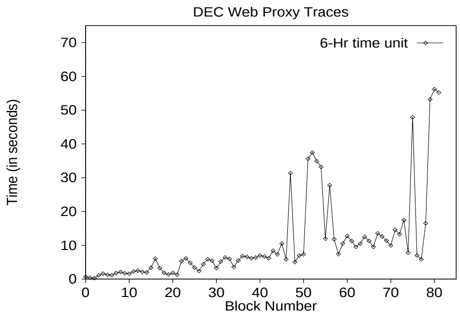


Figure 10: Time for Pattern Computation

| | Model Maintenance | Pattern Detection |
|---|---|---|
| MRW | $\surd$ | $\surd$ |
| UW | $\surd$ | $\surd$ |

Figure 11: Problem space enumeration in DEMON

experiments show that the notion of compact sequences discovered by our simple algorithm is meaningful and interesting. Second, the experiments also show that the results of our techniques are no panacea and need some post-processing and interpretation.

Figure 10 shows the time taken to incrementally update the set of existing compact sequences with a new block. (We numbered the blocks from $0$ to $81$, corresponding to the eighty two $6$-hour periods from noon of $9-2-1996$ to midnight of $9-22-1996$.) The spikes correspond to blocks which are significantly different from a large proportion of earlier blocks, since computation of the deviation between two significantly different blocks takes much longer than computation of the deviation between two similar blocks. (In the former case, both blocks are almost always scanned, whereas in the latter, they are scanned only rarely.) Not surprisingly, the block numbers corresponding to spikes fall into weekends, where the data characteristics are very different from the data characteristics of most weekday blocks.

# 6 Related Work

We first discuss incremental mining algorithms for frequent itemsets, clustering, and classification. In general, all algorithms we discuss now are designed for arbitrary insertions and deletions of transactions and hence do not exploit systematic block evolution. Moreover, they do not consider and cannot maintain models for the most recent window option with respect to an arbitrary block selection sequence.

In Section 3, we discussed the BORDERS algorithm for incrementally maintaining frequent itemsets. The FUP algorithm and its derivatives [CHNW96, CLK97, CVB96] are the first to address the problem of incrementally maintaining frequent itemsets. It makes several iterations and in each iteration, it scans the entire database (including the new block and the old dataset). The BORDERS algorithm improves the FUP algorithm by reducing the number of scans of the old database. Ester et al. [EKS$^+$98] extended DBScan [EKX95] to develop a scalable incremental clustering algorithm. In prior work, we developed a scalable incremental algorithm for maintaining decision tree classifiers [GGRL99b]. Utgoff et al. [Utg88] developed ID5, an incremental version of ID3, which assumes that the entire dataset fits in main memory and hence is not scalable.

Ramaswamy et al. [RMS98] segment the database of transactions into a sequence of time units to discover association rules that follow a user-defined pattern over these segments. They introduce the notion of a *calendar* to allow users to express interesting patterns. A calendar is a sequence of (possibly overlapping) time intervals. An association rule is said to *belong* to a calendar if the rule has the minimum support and the minimum confidence on each segment corresponding to a time unit in the calendar. Given a set of calendars, they discover all association rules that belong to the set of calendars. Our work differs from that of Ramaswamy et al. [RMS98] in two important aspects. First, they assume that the database is static and then discover association rules that belong to a calendar, whereas we maintain association rules as the database evolves. Second, each time unit of the database in the calendar is mined for association rules whereas we mine for a single combined model (belonging to one of several classes of models) over the set of selected time units.

Counting frequencies of itemsets using TID-lists was first proposed by Zaki et al. [ZPOL97]. They observe that it is too expensive to count using TID-lists the frequencies of all 2-itemsets. Our results explain this observation. Later, Sarawagi et al. also explored the use of TID-lists to count frequencies of itemsets [STA98]. However, they use TID-lists to count frequencies of all candidate itemsets in each pass. Overall, they observed that it is better to use a hash-tree (or prefix tree) instead of TID-lists. Again, our results explain the poor performance they observed: if the number of candidate itemsets is very high then PT-Scan outperforms TID-lists. Concurrent with our work, Dunkel et al. found that TID-lists are efficient for mining association rules on a special class of datasets which have a much higher number of items than the number of transactions [DS99]. In contrast, we look at incremental maintenance of association rules for any general transactional database.

## 7 Conclusions and Future Work

Figure 11 summarizes our contributions. We explored the problem space of systematic data evolution for two important objectives, model maintenance and pattern detection, and described efficient algorithms for both objectives. In future work, we intend to (1) explore the impact of the block granularity on the types of patterns discovered, and (2) to develop techniques to automatically determine appropriate levels of granularity.

## References


[AGGR98] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1998.

[AMS+96] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast Discovery of Association Rules. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 12, pages 307–328. AAAI/MIT Press, 1996.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In *Proc. of the 20th Int'l Conference on Very Large Databases*, Santiago, Chile, September 1994.

[CD97] S. Chaudhuri and U. Dayal. An overview of data warehouse and olap technology. *ACM SIGMOD Record*, March 1997.

[CHNW96] D. Cheung, J. Han, V. Ng, and C. Wong. Maintenance of discovered association rules in large databases: An incremental updating technique. In *Proceedings of the twelfth international conference on data engineering (ICDE)*, February 1996.

[CLK97] D. Cheung, S. Lee, and B. Kao. A general incremental technique for maintaining discovered association rules. In *Proceedings of the fifth DASFAA Conference*, April 1997.

[CVB96] D. Cheung, T. Vincent, and W. Benjamin. Maintenance of discovered knowledge: A case in multi-level association rules. In *Proceedings of the second international conference on knowledge discovery in databases*, August 1996.

[DH73] Richard Duda and Peter Hart. *Pattern Classification and Scene analysis.* Wiley, 1973.

[DS99] Brian Dunkel and Nandit Soparkar. Data organization for efficient mining. In *Proceedings of the 15th International Conference on Data Engineering*, pages 522–529, March 1999.

[EKS+98] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Michael Wimmer, and Xiowei Xu. Incremental clustering for mining in a data warehousing environment. In *Proceedings of the 24th International Conference on Very Large Databases*, pages 323–333, August 1998.

[EKX95] Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. A database interface for clustering in large spatial databases. In *Proc. of the 1st Int'l Conference on Knowledge Discovery in Databases and Data Mining*, Montreal, Canada, August 1995.

[FAAM97] Ronen Feldman, Yonatan Aumann, Amihood Amir, and Heikki Mannila. Efficient algorithms for discovering frequent sets in incremental databases. *Workshop on Research issues on Data Mining and Knowledge Discovery*, 1997.

[FPSSU96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

[GGRL99a] Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, and Wei-Yin Loh. A framework for measuring changes in data characteristics. In *Proceedings of the 18th Symposium on Principles of Database Systems*, 1999.

[GGRL99b] Johannes Gehrke, Venkatesh Ganti, Raghu Ramakrishnan, and Wei-Yin Loh. BOAT–optimistic decision tree construction. In *Proceedings of the ACM SIGMOD International Conference on Managment of Data*, June 1999.

[GRS98] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, June 1998.

[Gup97] Himanshu Gupta. Selection of views to materialize in a data warehouse. In *Proceedings of the International Conference on Database Theory*, January 1997.

[JD88] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[K.90] Fukunaga K. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, CA, 1990.

[Mue95] Andreas Mueller. Fast sequential and parallel algorithms for association rule mining: A comparison. Technical report, University of Maryland, August 1995.

[PH00] Vikram Pudi and Jayant Haritsa. Incremental mining of association rules. Technical report, DSL, Indian Institute of Science, Bangalore, 2000.

[RMS98] Sridhar Ramaswamy, Sameer Mahajan, and Avi Silbershatz. On the discovery of interesting patterns in association rules. In *Proceedings of the 24th International Conference on Very Large Databases*, pages 368–379, August 1998.

[SCZ98] Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *Proceedings of the 24th International Conference on Very Large Databases*, pages 428–439, New York, New York, August 1998. Morgan Kaufmann.

[STA98] Sunita Sarawagi, Shiby Thomas, and Rakesh Agrawal. Integrating mining with relational databases: Alternatives and implications. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 343–354, June 1998.

[TBAR97] Shiby Thomas, Sreenath Bodagala, Khaled Alsabti, and Sanjay Ranka. An efficient algorithm for the incremental updation of association rules in large databases. In *Proceedings of 3rd International Conference on Knowledge Discovery in Databases*, 1997.

[TK] Jeff Mogul Tom Kroeger, Carlos Maltazhn. Digital's web proxy traces. ftp://ftp.digital.com/pub/DEC/traces/proxy/webtraces.html.

[Utg88] P.E. Utgoff. ID5: An incremental ID3. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 107–120. Morgan Kaufmann, 1988.

[Wil88] P. Willett. Recent trends in hierarchical document clustering: A critical review. *Information Processing and management*, 24(5):577–597, 1988.

[ZPOL97] M.J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proceedings of the third international conference on knowledge discovery in databases and data mining*, 1997.

[ZRL96] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: An efficient data clustering method for very large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, June 1996.

*Venkatesh Ganti* is a PhD student at the University of Wisconsin-Madison. His research interests are in the areas of database systems and data mining (including online analytical processing, approximate query answering, and analyzing evolving data). He is a Microsoft Research Fellow. He received an MS from the University of Wisconsin-Madison and will receive a PhD in August, 2000.

*Johannes Gehrke* is an assistant professor in the Department of Computer Science at Cornell University. Gehrke's research interests are in database systems and data mining. He leads the Himalaya Data Mining Project and the Cougar Device Database System Project at Cornell University. Gehrke is the recipient of an IBM Faculty Partnership Award. He is the co-author of the textbook ”Database Management Systems (Second Edition)” published by McGraw Hill, and he holds two patents in the area of data mining.

*Raghu Ramakrishnan* is Professor of Computer Sciences and Vilas Associate at University of Wisconsin-Madison, and a founder and CTO of QUIQ, a company that powers online communities. His research interests are in the areas of database query languages (including logic-based languages, languages for non-traditional data such as sequences and images, and data mining applications), data visualization, and data integration. He is the recipient of a Packard Foundation fellowship and an NSF PYI award, and is on the editorial boards of The AI Review, Constraints, JIIS and JLP. He has also served on the program committees of several conferences in the database and logic programming areas, is Program Chair for KDD 2000, and is the author of the text Database Management Systems.