Incremental Updates of Inverted Lists for Text Document Retrieval *

Stanford University Technical Note Number STAN-CS-TN-93-1

Anthony Tomasic, Stanford University[†]
Hector Garcia-Molina, Stanford University[‡]
Kurt Shoens, IBM Almaden[§]

November 18, 1993

Abstract

With the proliferation of the world's "information highways" a renewed interest in efficient document indexing techniques has come about. In this paper, the problem of incremental updates of inverted lists is addressed using a new dual-structure index data structure. The index dynamically separates long and short inverted lists and optimizes the retrieval, update, and storage of each type of list. To study the behavior of the index, a space of engineering trade-offs which range from optimizing update time to optimizing query performance is described. We quantitatively explore this space by using actual data and hardware in combination with a simulation of an information retrieval system. We then describe the best algorithm for a variety of criteria.

1 Introduction

As the world's "information highways" proliferate and grow in capacity, they are providing access to an ever growing number of electronic document repositories. At each repository, the number of documents available on-line is rapidly increasing. At the same time, the number of end-users with network access is rapidly growing, and a variety of tools [8] such as World Wide Web and WAIS make it possible to reach even more information sources. This rapidly growing number of documents, sites, and user queries has brought about renewed interest in efficient document indexing techniques.

The underlying index structure for most document retrieval systems is the *inverted list* [7]. The inverted list for a particular word w contains a sequence of *postings*, each reporting the occurrence of w in a document. Each posting may include a variety of information, such as the word offset (within the document) where w occurs or the region where w occurs (title, abstract, author list, etc.) In a full text index, every word occurring in documents (minus perhaps some *stop* words) has

*This research was sponsored by the Advanced Research Projects Agency (ARPA) of the Department of Defense under Grant No. MDA972-92-J-1029 with the Corporation for National Research Initiatives (CNRI). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of ARPA, the U. S. Government or CNRI.

[†]Department of Computer Science, Stanford, CA 94305-2140. e-mail: tomasic@cs.stanford.edu

[‡]Department of Computer Science, Stanford, CA 94305-2140. e-mail: hector@cs.stanford.edu

[§]IBM Almaden Research Center. e-mail: shoens@almaden.ibm.com

an inverted list. As a rule of thumb, the size of the inverted lists for a full text index is roughly the same size as the text document database itself. In an *abstracts* index, only words appearing in the bibliographic information (e.g., title, abstract) have lists.

In an information retrieval system, users submit queries that consist of a set of words and some condition. The exact form of the condition varies: in a boolean system, queries are boolean expressions such as "(cat and dog) or mouse." In this example, the system would retrieve the inverted list for "cat" and "dog", intersect them, and then would union the result with the list for "mouse." The query may also give additional conditions, such as requiring that "cat" and "dog" occur within so many words of each other, or that "mouse" occur within a title region. In a vector model system, the query specifies weights for the words, and the system must locate documents that maximize the weighted sum of occurring words. Vector model systems typically use inverted lists to prune the set of candidate documents before the vector condition is evaluated.

Traditional information retrieval systems, of the type used by libraries (e.g., Stanford University's Socrates or the University of California's MELVYL) or information vendors (e.g., Dialog Inc. or Mead Data Central Inc.), assume a relatively static body of documents. Given a body of documents, these systems build the inverted list index from scratch, laying out each list sequentially and contiguously to others on disk (with no gaps). (They also built a B-tree that maps each word to the locations of its list on disk.) Periodically, e.g., every weekend, new documents would be added to the database and a brand new index would be built. Rebuilding the index is a massive operation, but its cost is amortized over multiple days of operation.

In many of today's environments, such full index reconstruction is not feasible. One reason is that text document databases are more dynamic. For instance, if one is indexing news articles, electronic mail, or stock information, the latest information is required. Thus, one would like to update the index in place, as new documents arrive. (Updating the index for each *individual* arriving document is inefficient, as we will discuss later. Instead, the goal is to batch together small numbers of documents for each in-place index update. To maintain access to the batch, it can be searched simultaneously with the larger index.)

A second reason why in-place updates are desirable is that they eliminate (or at least postpone) resource consuming reorganizations. Massive reorganizations may be acceptable in conventional systems where user load is minimal over weekends, but in today's world of 7 days a week, 24 hours a day continuous operation, degradation of service for prolonged periods is not acceptable.

A third reason why in-place updates may be desirable is that the index may simply be too massive for reorganization. As the volume of documents grows in some applications, it may be more desirable to have a dynamic index that can grow and dynamically migrate to new disk drives, without ever being fully reorganized.

In spite of the natural attractiveness of in-place index updates, very little is known about their implementation options or their performance. Systems that implement in-place updates typically use (as far as we know) relatively naive strategies that may be inefficient. For example, any time a WAIS index needs to grow an inverted list, it copies the whole list to a new disk area, leaving no free space at the end for future updates. Perhaps it would be more effective to leave some space, and to make additions that fit in that space? If multiple disks are available, can we stripe large lists across multiple disks to improve performance? Inverted lists vary tremendously in size: the ones for frequently occurring words can be huge, but there may be many that have only a few postings. What is the most effective layout of these lists to make their updates efficient? Which layouts lead to less disk space utilization? To better query performance?

Although we will not fully answer all these questions, in this paper we make the following contributions:

- A new dynamic dual-structure for inverted lists. Lists are initially stored in a "short list" data structure; as they grow they migrate to a "long list" data structure. Our proposed algorithm dynamically selects lists to migrate.
- A family of disk allocation policies for long lists. Each policy dictates, among other things, where to find space for a growing list, whether to try to grow a list in place or to migrate all or parts of it, how much free space to leave at the end of a list, and how to partition a list across disks.
- A detailed performance evaluation of the dual-structure lists and the various allocation policies. The evaluation is based on a collection of 64 days worth of NetNews that are indexed according to our algorithms. Our experimental system generates the exact sequence of disk block updates that each policy produces; this sequence is then executed on an IBM Risc System 6000 Model 350 computer with 3 disks to measure the update time. Based on the resulting disk layout, we also compute disk space utilization and estimate query performance. Real disk I/O operations have the advantage of not simplifying the dynamic nature of the execution (which occurs with a simulation).

In this paper we do not consider issues related to fault tolerance. We assume that the hardware is highly reliable. However, to be fair in estimating and comparing the I/O costs of various policies, we will periodically flush to disk all the data and directory information for each policy. In addition, the algorithms and data structures are constructed so that the incremental update of the index can be restarted if it is aborted. We believe fault tolerance issues are a rich area for future research.

In the next section we describe the dynamic dual-structure for inverted lists. In Section 3 we describe a model for the various allocation policies of inverted lists that reside on disk. We evaluate these policies in an experimental design described in Section 4 and report the results of the evaluation in Section 5. We then extrapolate our results to larger synthetic text document databases and describe the results in Section 6. In Section 7 we describe related work in the field. Finally, Section 8 concludes the paper.

2 Dual-Structure Index

In this paper we assume that when a new document arrives it is parsed and its words are inserted into an in-memory inverted index. At some point the in-memory inverted index must be written to disk. Our objective is to incrementally update the disk with the in-memory inverted index as efficiently as possible.

The lengths of the inverted lists for a database of text documents have a roughly exponential distribution (the Zipf curve [11]). This presents a dilemma for the in-place update of inverted lists since some inverted lists (corresponding to frequently appearing words) will expand rapidly with the arrival of new documents while others (corresponding to infrequently appearing words) will expand slowly or not at all. In addition, new documents will contain previously unseen words. Table 1 shows some statistical properties of a database of NEWS articles (cf. Section 4 for a complete

Text Document Database	News
Total Raw Text	686 MB
Total Words	788,256
Total Postings	48,526,577
Documents	138,578
Average Postings per Word	61
Frequent Words	39,413
Infrequent Words	748,843
Postings for Frequent Words	93.6%
Postings for Infrequent Words	6.4%

Table 1: Statistics for a News abstracts text database. Abstracts databases index general information about a document such as author names, title, the set of words in the abstract, etc. A frequent word for this table ranks in the top 5% of all words (in order of frequency). Postings for frequent words are given as the percentage of all postings in the database. Infrequent words are all words that are not frequent.

description of the database). For example, if we consider frequently appearing words as those that rank in the top 5.0% of all words (in order of frequency) we see that the postings for the frequently appearing words account for the vast majority (93.6%) of the postings.

In our scheme there are two data structures for lists. We place short inverted lists (of infrequently appearing words) in fixed size blocks where each block contains postings for multiple words. These lists are referred to as *short lists* and the fixed size blocks are known as *buckets*. The idea is that every inverted list starts off as a short list; when it gets "too big" (see below) it becomes a long list. We place the long inverted lists (of frequently appearing words) in variable length contiguous sequences of blocks on disk. We refer to these inverted lists as *long lists*. Each block of a long list contains postings for only one word. Given a word w, we examine a *directory* which determines if the word has a long inverted list. If the word does not have a long inverted list, it has a short inverted list or no inverted list at all. In this case, a function $h(w)$ (e.g., a hash function or a tree search) returns the bucket where the short inverted list, if any, for the word is stored.

At some point, an in-memory list L for word w (generated from arriving documents) must be moved to disk. First, if w already has a long list (on disk), L is appended to the long list as discussed in the next section. Otherwise, we assume L is a short list and insert it into bucket $h(w)$. If the bucket is not already in memory, it is read in, and L inserted. (If a list for w already existed in the bucket, L is added to it; else a new short list is created in the bucket.) If the bucket overflows, we then pick the longest short list in block, say M, remove it, and make M a long list. Once M is removed, the bucket will be partially empty. The updated bucket $h(w)$ is written out (eventually), and list M is written out to disk as discussed in the next section. Note that a word w never has both a short list and a long list associated with it. The buckets serve as a *filter* for words with inverted lists containing only a few postings, since these words are unlikely to grow enough to merit migration into a long list. (Assuming that the bucket data structure is large enough to hold all the infrequent words.)

Figure 1 shows an animation of the behavior of buckets. We choose bucket 0 as an example bucket and run the bucket algorithm for a short time on a small system. Data is from a News, as

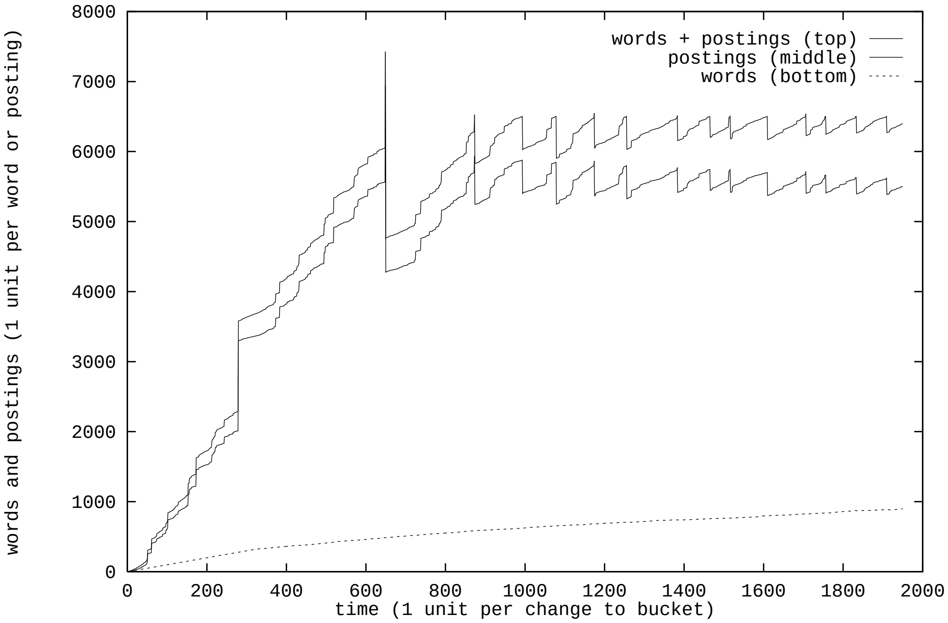

Figure 1: An animation of the behavior of bucket 0 for the first 6 updates for a system with 250 buckets. The top line is words + postings, the middle line is postings, and the bottom line is words.

explained in Section 4. The bucket has a size of 6500 units, where each posting is charged 1 unit and each word is charged one unit too. (For each inverted list in the bucket, we need to store the word it represents plus all of its postings.) Each time step on the x-axis corresponds to a change in the bucket – the insertion of a new word with its postings, the appending of postings to an existing word, or the removal of a word and its postings from the bucket. The y-axis measures the combined number of words and postings in the bucket. The top line in the figure is the total number of words and postings, the middle line is the total number of postings and the bottom line is the total number of words in the bucket. For the total number of words in the bucket, we see a slow rise in the number of words in the bucket as new words are continuously inserted into the bucket. For the number of postings in the bucket (the middle line), we see a steep climb as the bucket fills up and two leaps where a very large in-memory list is inserted into the bucket at approximately time 300 and time 700. The second insertion of the large in-memory list causes an overflow and the list is removed from the bucket, shown as a downward spike in the graph. After the spike, the bucket continues to fill to about time 900 where it overflows and again the largest short list is removed from the bucket. This behavior is repeated for the rest of the time on the graph. The top line in the graph, showing the total of words and postings, stablize at 6500 units since the overflow spikes in the graph always occur at the size of the bucket. Note that the size of the spikes get smaller over time. Since the set of frequently appearing words is constant over time, once a frequently appear word overflows into a long list, it never interacts with the bucket data structure again. This leaves only shorter inverted lists as the longest list in any bucket.

In summary, the dual-structure index allows us to apply different storage structures to the huge number of infrequent words and to the relatively few frequent words. Through the use of fixed-size buckets, this approach dynamically discovers the frequent words that require their own long list.

Updates to the large number of infrequent words are amortized into a relatively small number of disk operations, since the buckets are small enough to fit in memory. In addition, coalescing infrequent words reduces wasted disk space due to allocation of complete disk blocks to very short lists.

3 Policies for Allocation of Long Lists

In this section we present policies for the allocation of long lists to disk. Implementations of IR systems indexes merge inverted lists to compute the answer to a boolean query. This is possible because the document identifiers appear in sorted order in inverted lists. We assume that new documents are numbered with identifiers in increasing order and that all long lists are updated by appending new postings to them. With these assumptions, the merge operation can be used to compute answers to boolean queries with our long list data structure.

Long lists are created initially by the overflow of a bucket. Once a word has a long list on disk, subsequent in-memory lists for that word will be appended to that long list. In this section we consider only long lists and refer to them by the shorthand *lists*.

In allocating lists to disk, there are two extremes policies. One extreme policy optimizes the time to incrementally update (the *update time*). Let L be the list for a word w and let M be the in-memory list to be appended to L. If M is written sequentially at the current end of the data on disk, irrespective of L, then update time is optimized because the disk head never seeks during a sequence of updates. The other extreme policy optimizes the reading of a list during query processing (the *query time*). To update a word w, we can read L from disk, append M to it, and write the new combined list to a new location on disk. This optimizes query time because exactly one seek is required to read any list. However, the query time for the update optimized policy is poor (since the list for a word will be spread over the disk) and the update time for the query optimized policy is poor (since reads are intermixed with writes). These disadvantages are magnified by the number of updates to a word. Thus, if n updates have been processed under the update optimized policy, n reads are required to read an entire list for a word since a new part of a list is generated for each update. For the query optimized policy, the portion of the list written on the first update must be read and written in each of the subsequent $n-1$ updates since it is copied for each update.

Consider an additional technique as a compromise. When initially writing a list L for a word w, we reserve additional blocks at the end of the list. Then, when M is to be appended to L, we check if M fits in L's reserved space. If it does fit, then we append M to L with an in-place update. If M does not fit, we can resort to one of the extreme policies above.

For the above policies, parts of the list for a word are allocated in contiguous regions of disk called *extents*. To distinguished variable sized extents from fixed sized extents, we use the term *chunk* for variable sized extents and reserve the term extent for fixed sized contiguous regions of disk. (We also study the extent case later.) Multiple chunks for an inverted list may be allocated. The pointers to all chunks are recorded in the directory. The directory entries for a word may point to chunks on multiple disks. The directory resides in memory at all times. Periodically, the directory is written to disk.

Given the above discussion, we describe policies as consisting of three orthogonal strategies:

1. Strategies to assign a disk unit to a new word or chunk,

Variable	Value	Meaning
Limit	0	Never update in-place
	z	Update in-place if enough space
Style	fill ($e = 3$)	Fill in fixed size extents
	new	Write a new chunk when appropriate
	whole	Long lists are single whole chunks
Alloc	constant ($k = 10$)	Constant extra postings reserved
	block ($k = 3$)	Multiple of a fixed sized block reserved
	proportional ($k = 1.1$)	Proportional extra postings reserved

Table 2: The variables and values that determine a policy for the allocation of long inverted lists to disks. The values in parenthesis are for each allocation strategy or style.

2. Strategies for appending in-memory lists to long lists, and
3. Strategies to allocate space on the disk.

There are various choices for each of the above strategies. By selecting a choice for each strategy, a policy is defined. We consider each strategy in turn. Obviously many more choices exist than what we will describe here. However, we believe that the broad terrain of choices available to the system designer is considered here.

3.1 Strategies to assign a new word or chunk to a disk

When the list for a new word w is added to the directory or a new chunk of a list for a word w is allocated, a disk is chosen. Let there be n disks, numbered from 0 to $n-1$, and let i be the disk chosen when the last new word or chunk was allocated (i is initially 0). The strategy considered here is to chose disk is $i + 1 \bmod n$. (Other strategies could be to look for the most empty disk or a disk where the list has the fewest chunks. These strategies are not considered in this paper to keep the space of possible solutions manageable.)

3.2 Strategies for combining in-memory and long lists

We have an in-memory list M that we wish to append to a long list L. Let x be the size (in postings) of the long list, let y be the size (in postings) of the in-memory list, and let z be the size (in postings) of the space remaining in the chunk which can accommodate new postings. As described in Section 3.3, when a chunk is allocated to disk, it may have *reserved space* at the end of the chunk where future postings may be append. Note that z may be zero or positive and that x and y are always positive. Our strategies for appending will call on the following basic operations which operate with respect to long list L. The RELEASE list is used to delay the deallocation of long lists while they are copied.

UPDATE(a) reads the last block containing postings, appends the in-memory a to it, and then writes the result back as an in-place update.

READ(a) reads all the postings for L, places L on the RELEASE list, and returns the postings read as in-memory list a.

```
1.  if $y \leq Limit$ then
2.     UPDATE($M$) update long list in-place with the in-memory list
3.  else
4.     if $Style = whole$ then
5.        READ($a$) read long list
6.        WRITE RESERVED($M$ and $a$) append in-memory list and write with reserved space
7.     if $Style = fill$ then
8.        WHILE ($M$ not empty) in-memory postings remain
9.                 WRITE($M, M$) write in-memory postings
10.    if $Style = new$ then
11.       WRITE RESERVED($M$) write in-memory postings with reserved space
```

Figure 2: The algorithm for updating long lists.

WRITE(a,b) writes up to e blocks worth of postings from a and returns the remaining postings as in-memory list b. The global parameter e is called the *extent size.* (The fill style, below, breaks up in-memory lists into extent size chunks.) If a contains less than e blocks worth of postings, e blocks are still allocated on disk.

WRITE RESERVED(a) writes the a in-memory list to disk with reserved space at the end of the list (cf. Section 3.3).

A strategy for appending an in-memory to a long lists is specified by two variables, *Limit* and *Style*. *Limit* is either 0 or z. *Style* is *fill*, *new*, or *whole*. Table 2 summaries the variables and values governing policies.

Figure 2 shows the algorithm for updating long lists. The first three lines check if the existing chunk can and should be extended with the in-memory postings. If this isn't possible or desirable ($Limit = 0$), the fourth through sixth lines (whole) copy the old postings to a new location with the in-memory postings appended. Lines seven, eight and nine (fill) write out multiple extents. Lines ten and eleven (new) write a new chunk with reserved space. Once consequence from lines one and two is that an in-memory inverted list is never split into two different chunks for an in-place update.

Periodically, the buckets and the directory are written to disk. At this time, the disk blocks for the previous buckets and directory are returned to free space for the disks. In addition, in the case of the whole strategy, the old long lists on the RELEASE list are returned to free space for the disks. Note that the WRITE RESERVED operation has three different ways of determining how much space must be reserved when a list is written to disk (described in the next section). However, the WRITE operation writes a fixed amount of space determined by the parameter e.

3.3 Strategies to allocate space on a disk

Given a request for a chunk of size f and a disk, we need a contiguous region of free space on the disk to satisfy the request (this is similar to the problem of free space allocation for blocks in a file system). We use a first-fit strategy by scanning the free list for the disk from the beginning of the disk. Upon finding a contiguous sequence of f or more blocks, the chunk is placed at the beginning of the free blocks and the remaining free blocks are returned to free space. (Other strategies could be to best-fit or to use a buddy system. These strategies are not considered in this paper to keep the space of possible solutions manageable.)

In addition, the WRITE RESERVED call reserves space at the end of every list for future growth. That is, additional space is allocated to a chunk to hold postings which will appear in subsequent updates. Let x be the size (in postings) of the inverted list being written to disk and let $f(x)$ be the allocated space (list plus reserved space). The resulting size (in blocks) of a chunk is the number of blocks needed to hold $f(x)$. For the new style x is typically the size of an in-memory list. For the whole style x is typically the size of the entire long list for a word.

We consider three choices for the definition of $f(x)$. The *constant* strategy adds a constant number k of postings to the end of the inverted list, i.e., $f(x) = x + k$. The *block* strategy insures the chunk is of constant multiple of size k, i.e., $f(x) = k \cdot \lceil \frac{x}{k} \rceil$. (In practice, we specify k for this strategy in terms of blocks instead of postings.) Finally, the *proportional* strategy allocates a chunk in proportion to the number of postings being written, i.e., $f(x) = kx$. The variable *Alloc* equals *constant*, *block*, or *proportional*, for the corresponding choice of strategy.

3.4 Policies

A *policy* is determined by the values of the variables *Style*, *Limit* and sometimes *Alloc*. If *Limit* = 0, then any reserved space for a chunk is never used, so we automatically set *Alloc* = *constant* with $k = 0$. If *Style* = *fill* then the allocation strategy is irrelevant since it is never considered.

Consider the update optimized policy described earlier. This can be achieved by setting *Limit* = 0 and *Style* = *new*. This policy minimizes update time by simply writing out the update list blocks as fast as possible. No reading is done because no updates in-place occur. We expect that this policy will have the best update time and that the query time for the resulting index will be poor.

Consider the policy for fast queries. Let *Limit* = 0 and *Style* = *whole*. Setting *Style* to *whole* insures that the inverted list for any word will always be a single contiguous chunk and thus query time is minimized. We expect that the update time for this organization will be high, due to the amount of moving of lists that must be done. To ameliorate this situation, we can let *Limit* = z and *Alloc* = *proportional* with a constant of, say, 1.1 (i.e., reserved space that is 10% of the size of the long list). With each move of the long list, the reserved space will grown by 10%, permitting more and more in-place updates of in-memory lists.

Finally, we consider a policy that attempts a trade-off: to minimize query time via disk striping and keep the cost of updates low by organizing inverted lists into chunks that never move once they are full. Let *Limit* = z and *Style* = *fill* with an extent e size, say, 3 (blocks). With this policy, each inverted list will grow until it reaches the limit of its chunk and then a new chunk will be started on a new disk. We expect comparatively good query and update times for this policy.

So far our discussion has focused on the addition of documents to an index since typically databases only grow in size, or deletion is infrequent enough that the entire index is rebuilt. The addition of incremental deletion of documents poses some problems to the design of an index. One method maintains an index of document identifiers and all the words in the document (or the words are extracted from the original document). Given this index, each inverted list for a word in the document would be fetched, the reference to the document deleted, and the new inverted list

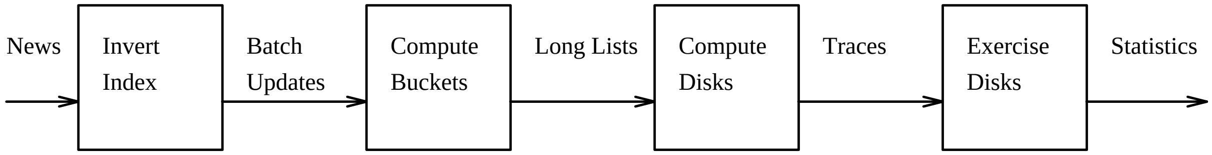

Figure 3: The flow of data for the experiment design. Arrows represent data. Boxes represent the transformation of data via a process.

rewritten to disk. However, the size of this index is the same as the size of the inverted index. To avoid this cost, existing implementations typically maintain a list of deleted document identifiers and filter any answer to a query through this list. This deletes the document from the point of view of the user since a deleted document identifier will never appear in an answer. To reclaim the space taken by the deleted document identifiers in the index, a background process sweeps the lists in the index one list at a time, removing any deleted documents. After a sweep of the index, the list of delete document identifiers can be thrown away. Since this issue is orthogonal to the the issues in this paper, we do not consider deletion further.

In summary, the parameters described in this section span an interesting range of approaches to storing long lists. By varying these parameters, we can model schemes that keep the lists sequential and those that break the lists into contiguous chunks. We can control the size of the chunks allocated, either as fixed-length chunks or as chunks whose size is controlled by the frequency of a word. Finally, we can control whether unused space at the end of a list is filled in or not. The choice of parameters permits trade-offs between index build performance, query performance, and index disk space consumption.

4 Experiment Design

Figure 3 diagrams the flow of data for our experiments in building inverted indexes. Each arrow represents a data set and each box represents a process that transforms data sets. The diagram also serves as an outline for this section and we describe each part of the diagram in turn.

4.1 News

The source text document database is 64 days of News articles gathered from November 18th, 1992, to January 21st, 1993. (December 10th is missing.) See Table 1 for statistics on the database. Once per day the the local server was scanned for new documents. News documents less than 2560 characters in length were eliminated to increase the average document size to a more typical range of about 5K characters [2]. In addition, to eliminate non-text documents, the frequency occurrence of the vowels in each document were measured for a 1K sequence of characters located either near the end of the document (for documents of 4K characters or less) or from an offset of 3K characters from the beginning of the document (for documents greater than 4K characters in length). If a frequency did not meet a corresponding threshold value, the document was eliminated on the assumption that the document contains non-English language text. An inspection of a random sample of the documents from the remaining database shows that the filter worked very well.

Each day of documents is a *batch* and is processed separately from other days. (We do not consider document deletions in this study.) While the dual-structure index does not require periodic

```
for years. And it was a total flop, in all the years it was available very
few people ever took advantage of it so it was dropped.
```

(a)

```
a advantage all and available dropped ever few flop for in it of
people so the took total very was years
```

(b)

Figure 4: (a) A fragment of a document from November 18th, 1992; (b) the tokens resulting from the document fragment.

abandons	1	abashed	2	abate	1	abated	1	abatement	1	abb	2

Table 3: A part of the batch update for November 18th, 1992, shown as pairs of words and the number of documents the word occurs in.

updates, this arrangement is good for measuring activity at periodic intervals.

4.2 Invert Index

The *invert index* process accepts a sequence of document batches as input, processes them, and generates a *batch update* for each batch. A batch update contains a list of words that appear in the documents of the batch and the number of times each word occurs in the batch. A word and its frequency of occurrence is termed a *word-occurrence pair.*

To generate a batch update, each document in the batch is lexically analyzed to produce a token stream. Sequences of letters and sequences of number are tokens – all other characters are ignored. Certain lines of a document (such as "Date: " lines) are also ignored. Finally, duplicate tokens for a document are dropped. After all documents for a batch are reduced to sets of tokens, an inverted file is constructed for the batch. Tokens are converted to words by converting upper case letters to lower case. The batch update containing all the words and the lengths of the inverted lists for each word is then constructed. Figure 4a shows a fragment of a document and Figure 4b shows the resulting set of tokens from the document fragment. Table 3 shows a part of a batch update. Note that the misspellings of words are part of the batch update as well. At this point all words in batch updates are converted to unique integers to simplify the remaining computations.

An implementation of an information retrieval system proceeds in the same way we have described here, except that it would keep, for each word, its complete inverted list, as opposed to the simple word-occurrence pair we keep here. For our performance evaluation, we do not need to know the contents of each inverted list, only its size, which is what the word-occurrence pair gives us. note, thus, that our batch update is our representation of the in-memory index of Section 2.

A second difference between the inverted index process and a real implementation is that a real system would drop some frequently occurring words from the in-memory index (via a stop list), and tokens would be reduced to their word stems via a stemming algorithm. Since stop lists and stemming algorithms vary widely, we have chosen not to use either. The data used here is a base line measurement which can be translated to a given implementation once the effects of a specific

Variable	Value	Description
Buckets	1,500	Number of buckets
BucketSize	6,500	Size of bucket in words and postings
BucketTotal	9,750,000	*Buckets* · *BucketSize*
BlockPosting	683	Postings per Block
Disks	3	Number of Disks
BlockSize	4,096	Bytes per Block
BufferBlock	400	Memory blocks for the I/O buffer

Table 4: The experimental parameters and base case values.

```
0 0
125144 4746
133663 4123
180761 4084
124376 3637
313269 4567
```

Figure 5: A part of the output of the compute buckets process. Each line is a word-occurrence pair. The left column contains integers representing words. (Words are numbered alphabetically.) The right column contains the lengths of the corresponding in-memory lists. The line "0 0" indicates the end of a batch update.

stemming algorithm and stop list on the vocabulary of a text document database are understood.

4.3 Compute Buckets

The *compute buckets* process takes the sequence of batch updates as inputs, runs the bucket algorithm described in Section 2 on the sequence (we use a modular arithmetic hash function for $h(w)$), and generates a single trace file of updates to long lists. Each update in the file indicates the word involved, and the number of postings to be add to the corresponding long list on disk. (Note that the postings for an update can comes from the new postings in a batch or from previous postings in a bucket.) In addition, a marker for the end of each batch update is added to the trace. Figure 5 shows a part of the output of the compute buckets process. For instance, in the second line of this figure, the number 125,144 is the unique identifier for a particular word and the number 4,746 is the number of postings to be appended to the long list for that word.

Table 4 lists the variables that control the bucket computation and the base values used for those variables in the experiments reported in the next section. Variable *Buckets* records the number of buckets and variable *BucketSize* records the size of each bucket (we count 1 for each word and posting placed in a bucket). Variable *BucketSize* implicitly models the efficiency of the compression algorithm applied to in-memory inverted lists since computations are in terms of postings instead of bytes. The remaining variables in this table we describe in the next section.

In comparing the compute bucket process and an implementation of the bucket data structure in an information retrieval system, we note that an implementation would perform a similar computation using inverted-lists as the compute bucket process does using word-occurrence pairs.

```
update bucket disk 0 id 0 size 1587
update bucket disk 1 id 0 size 1587
update bucket disk 2 id 0 size 1587
update chunk disk 0 id 0 size 0
write word 125144 posting 4746 disk 1 id 1587 size 7
write word 133663 posting 4123 disk 2 id 1587 size 7
write word 180761 posting 4084 disk 0 id 1587 size 6
write word 124376 posting 3637 disk 1 id 1594 size 6
write word 313269 posting 4567 disk 2 id 1594 size 7
```

Figure 6: An I/O trace corresponding to the previous figure.

A implementation would produce the same set of long lists. We assume that during the update process the buckets are kept in memory since they are referenced much more frequently than the long lists. At the end of each batch update, all buckets are flushed to disk. Note that the cost of maintaining all the buckets in memory during the update process can be avoided by sorting the in-memory lists into bucket order and then merging the in-memory list with the buckets, requiring only one bucket to be in memory at any single point in time.

4.4 Compute Disks

The *compute disks* process takes as input the trace file of long list updates and computes the sequence of I/O systems calls required to implement the policies described in Section 3. In addition, the write operations for saving the buckets and the directory are added at the end of each batch update. Figure 6 shows a sample of a I/O trace file. The first three lines indicate that the write of the bucket data structure occurs on three disks starting at location 0 and continuing for 1587 blocks. The next line writes an empty directory. (The directory is empty because this is the beginning of the trace, i.e., no long lists have been written – no actual I/O is performed for this line). The following lines write inverted lists for each word. For instance, the first "write word" line indicates that word 125,144 writes 4,746 postings on disk 1 starting at block 1,587 for a size of 7 blocks.

In addition to Table 2, Table 4 lists the variables and values used for the compute disks process. Note that the variables *BlockPosting* and *BlockSize* implicitly model the efficiency of the compression algorithm applied to long lists. A disk trace corresponds closely to the sequence of system I/O calls an implementation would perform for a given policy.

4.5 Exercise Disks

The *exercise disks* process takes a trace of I/O operations as input and executes it on an IBM RS 6000 Model 350 computer (64 MB memory, UNIX AIX 3.2 operating system with 3 disks (Seagate ST41200NM, 1 GB capacity, 5.25 inch, SCSI-1 standard) and an I/O bus (SCSI-1 standard). Each line of the trace generates a read or write system call request and after the update of the buckets and the directory all system buffers are flushed to disk.

Requests to each disk are issued by independent processes to achieve maximum parallelism. Request are directed to "raw" partitions of the disk, bypassing the operating system's file system and disk buffer pool. Our algorithms do not revisit the same blocks within a single batch, thus eliminating the advantage of a buffer pool. In addition, we assume that relevant data will not remain in buffers from one batch update to the next. Furthermore, bypassing the file system saves CPU overhead and results in slightly superior data rates. Finally, bypassing the operating system isolates our experiment design from effects introduced by the file system and thus experiments are independent of any particular file system implementation.

One drawback to using raw disk partitions is that the operating system obeys the disk requests exactly and does not coalesce adjacent write requests into single disk I/O operation. For this reason, the disk exerciser program does its own coalescing of I/O operations where possible without reordering the execution trace. To be faithful to real systems with a finite amount of buffering, the disk exerciser will only coalesce up to *BufferBlock* blocks (each of size *BlockSize*) in a single request.

4.6 Discussion

The experimental design presented here has many advantages. One of the most important is the decoupling of each process from the subsequent process, which permits varying parameters of a process to study the effects on the corresponding data transformation. However, the design rests on the assumption that the CPU costs of each process do not dominate the total computation time.

An in actual running IR system, processing is divided into two phases. In Phase I, documents must be parsed and an in-memory inverted index built. In this paper we are not address Phase I processing times, since they are independent of the disk structures used to represent the index and of the incremental index update policies.

In Phase II, the in-memory index is merged with the disk index. We wish to study the total time for this. In a real system, Phase II consists of CPU operations (e.g., appending an in-memory list to a copy of a long list read from disk) and I/O operations (e.g., reading and writing the long lists). The CPU and I/O operations are typically overlapped. We believe that the Phase II CPU time is small relative to the total I/O time, so that due to the overlap, the total time will be dominated by the I/O time.

Thus, it is sufficient to measure the time to execute the I/O operations to estimate the total time of a Phase II update. This is what an our experimental system does: it generates *exactly* the same stream of I/O operations a real system would (although the *contents* differ), and measures the time to execute the stream.

To test our belief that Phase II CPU costs can be overlapped with I/O operations, we need to test an actual running IR system. We selected the Rufus system [9] and built an inverted index for 307 megabytes of documents from a collection of IBM internal bulletin board articles. First the Phase I requirements were measured at about 9 minutes and then the real Phase II time required to build the index in two situations was measured. The first situation constructed the index on disk for the parsed documents. The second situation performed the same computations for the parsed documents, except that the I/O operations were not performed. The first situation required 199 minutes. The second situation required 45 minutes. Thus, in the worst case with no overlap of CPU and I/O operations, about 77% of the time is spent to build an index in I/O operations. We believe that this sufficiently justifies our experimental design since CPU utilization is probably

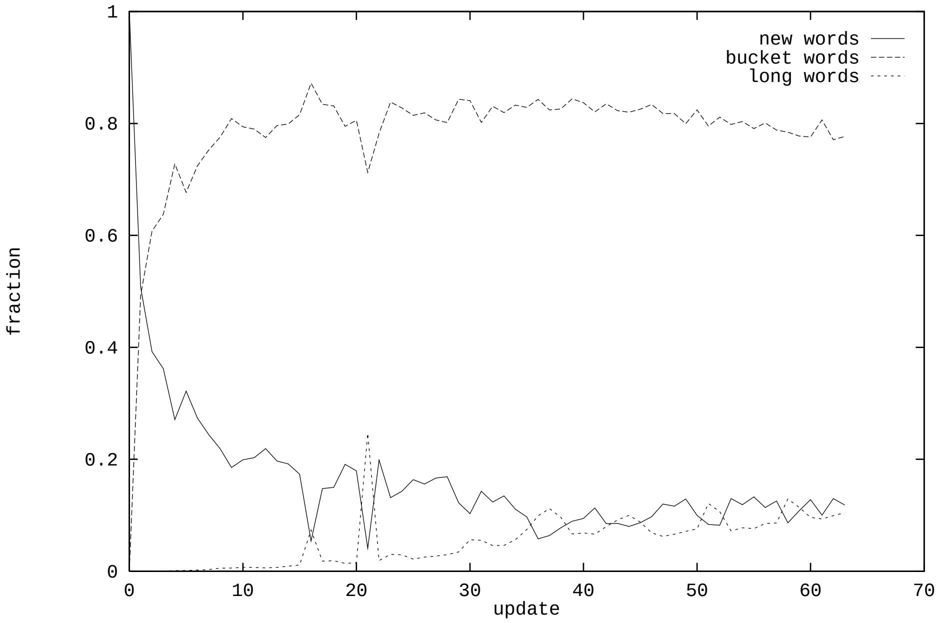

Figure 7: The fraction of words per update in each category.

invariant to the database indexed. However, we caution the reader that the real times presented in Section 5.3 do not include the Phase I times to build an inverted index.

5 Results

In this section we describe results for a set of experiments. An *experiment* is the execution of the sequence of processes described in Section 4 over the 64 days of collected data[1] using the defaults values given in Table 2 and Table 4, except as otherwise noted. The values of variables are sometimes systematically varied to show the sensitivity of a variable to some measurement. An *update* refers to the incremental batch update of the index. Some measurements apply only to the update. Other measurements apply to the index which results from the sequence of updates. In this case we refer to the *index after update*. For this section, the *final index* is the index which is produced after all the updates have been processed. We describe the results for each process in turn.

5.1 Compute Buckets

In this section we show the behavior of buckets and the generation of long lists. In addition, we tune the number of buckets and the size of a bucket in the bucket data structure.

Bucket behavior To show the behavior of the buckets, we measure the number of long lists in an update. For each word-occurrence pair in an update, we can categorize the word of the pair as

[1] We explore larger data sets in Section 6.

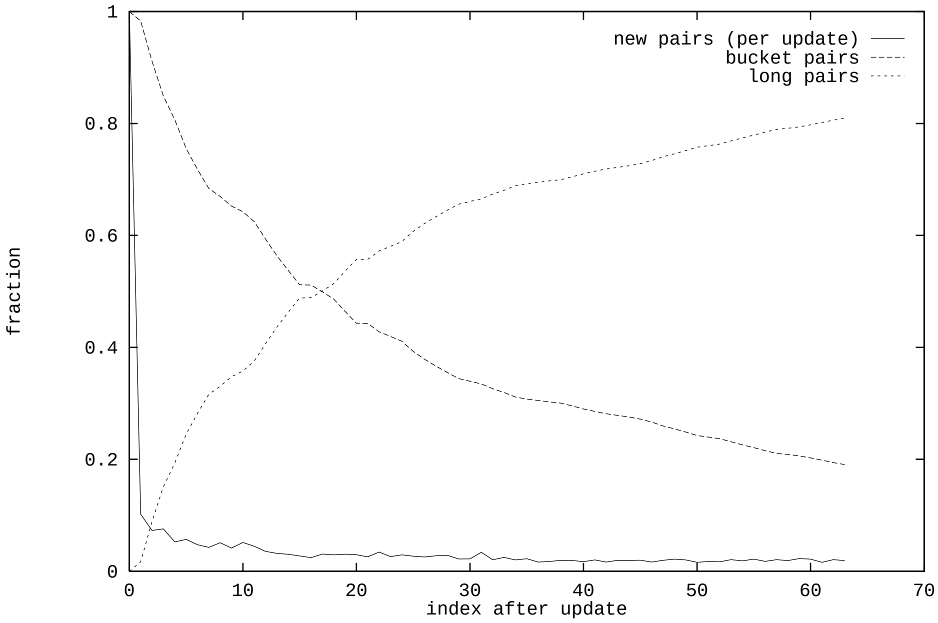

Figure 8: The fraction of bucket words and postings and long words and postings in the index. In addition, the new pairs curve shows the per update fraction of the index from new words.

one of three types: a *new* word (a previously unseen word), a *bucket* word (a word that is already in a bucket), or a *long* word (a word that has a long list).

Figure 7 shows, for each update, the fraction of words belonging to each category. Observing the "new words" curve, we see that initially all word-occurrence pairs contain new words since the buckets are empty and there are no long lists. This behavior very rapidly drops off as the buckets fill up with frequently appearing words. Eventually, after about 40 updates, the fraction of new words per update stabilizes around 10%. Observing the "bucket words" curve, we see that the fraction of words in an update which are in buckets rapidly rises until about the 15th update. Since the majority of words are the same in every update, this rise indicates that the buckets are filling up with words. The curve declines (roughly linearly) after update 40 as new words (containing typically short in-memory lists) fill the buckets and cause words to overflow into long lists. Observing the "long words" curve, we see that initially, no word-occurrence pairs contain long words because a few initial updates fit into the bucket data structure. The fraction of long lists rises (roughly linearly) after the buckets fill up in the initial stage. (The spike at update 21 is due to a very small size of the update for that day introduced by an interruption in the gathering of data.) Finally, we note a periodic set of peaks spaced seven days apart on the "long words" curve. Each peak corresponds to a Saturday when the corresponding update is smallest for the week. Small updates have higher fractions of frequently appearing words.

Another measure of the behavior of buckets is the flow of postings from the bucket data structure to long lists. At the end of each update, the total number of words and postings in the bucket data structure and the total number of words and postings in long lists are counted, as shown in Figure 8. The "bucket pairs" curve shows that the fraction of the index in buckets steadily drops as (a) words overflow and acquire long lists and (b) as long lists are appended with addition postings.

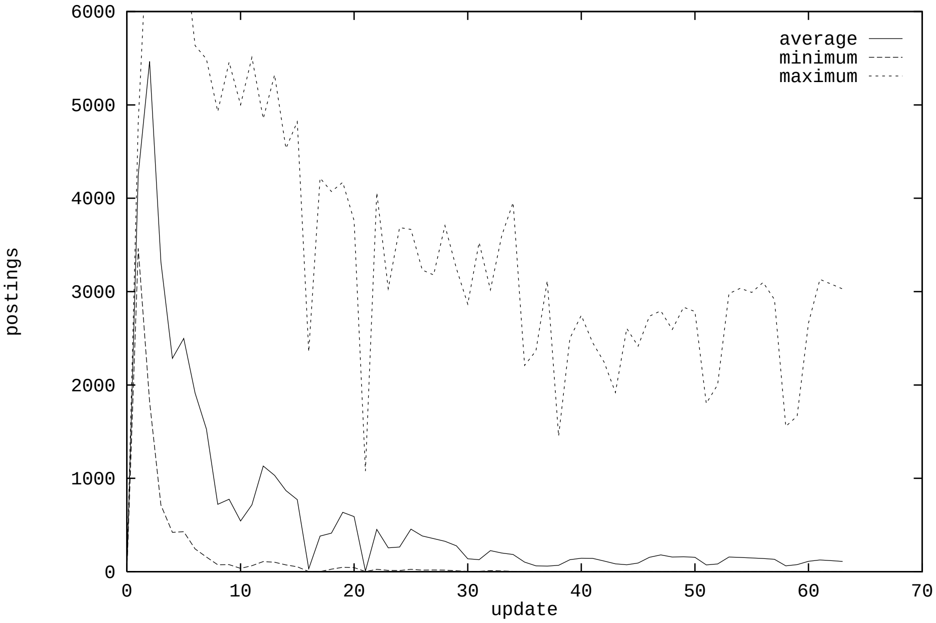

Figure 9: The average, minimum and maximum length of a long list update. The x-axis is the update number. The y-axis is in postings and it cuts off four points on the curve for the maximum number of postings per update. The first seven points of the maximum curve are (0, 0), (1, 4816), (2, 7329), (3, 7488), (4, 6380), (5, 6732), (6, 5635).

The "long pairs" curve shows that the fraction of the index in long lists steadily climbs for exactly the same reasons. The corresponding points on the two curves total to 1.0 – the entire index – since every word and posting must be in one of these two categories at the end of an update. In addition, the "new pairs" curve shows the fraction *per update* of new words and postings with respect to the entire index quickly stabilizes around 2% of the index. (The curve starts at 1.0 because on the first update, all words and postings are new.) The constant influx of new words and postings, which typically have small in-memory lists, implies a steady overflow of words from buckets and thus implies a steadily increasing number of words with long lists.

Another principle measure of the output of the bucket data structure combines the number of long lists with the number of postings. Figure 9 shows the maximum, average, and minimum number of postings in a long list update (produced by a bucket overflow or by an in-memory list) for each update. The curve starts at zero since the initial update fits in the bucket data structure. The curve then spikes sharply to over 7,000 postings as a few very long lists overflow from buckets. The long lists are the product of several updates. The average then quickly declines through the 10th to 30th updates as the number of long lists rises and shorter bucket lists overflow. The performance of the compute disk process is directly proportional to the number of postings which must be stored. Thus, we expect that performance declines as updates are incrementally applied to the index since the average number of postings per long list update declines. In Section 6.3 we accommodate this behavior by expanding the bucket data structure.

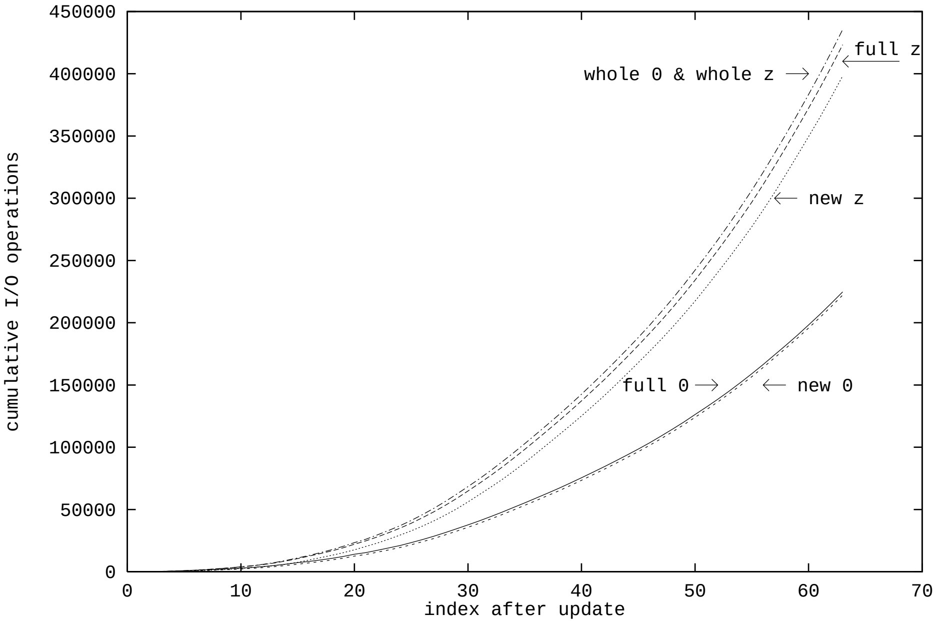

Figure 10: The *cumulative* number of I/O operations needed to build the final index. The x-axis is the index after the given update. The y-axis is the cumulative number of I/O operations needed to incrementally build the index. Each curve is label with the values of *Style* and *Limit*.

5.2 Compute Disks

In this section we consider the effects of the various allocation strategies described in Section 3. As our unit of measurement for this section only we count I/O operations. Each I/O operation corresponds to a call to the operating system that results in a disk seek and transfer of information. Note however that counting I/O operations is only an estimate of the time taken for a sequence of I/O operations. We consider the actual time taken for I/O operations when the exercise disk results are presented in Section 5.3. We study I/O operations in addition to actual times because they several insights into the behavior of the long list policies, a wider range of parameters can be studied simply because each experiment takes less time, and I/O operations closely estimate actual times. To compare allocation strategies, first we compare the three styles with zero reserved space to study the effect of in-place updates with respect to index build time, disk space utilization and the query performance of the resulting long lists. Then allocation strategies are added to study the effects of reserved space for the same issues.

5.2.1 Styles

The number of I/O operations needed for each of the three policies is shown in Figure 10. In the case that *Limit* = z, we use *Alloc* = *constant* with a constant of 0. This removes the effect of the allocation policies. The x-axis is the index after the given update. The y-axis is the *cumulative* number of I/O operations needed to incrementally build the index. The first observation is that all the curves in the graph have increasing slope. This means that the time to run each update takes longer as the index grows in size. This is due to the increasing number of long lists. Second,

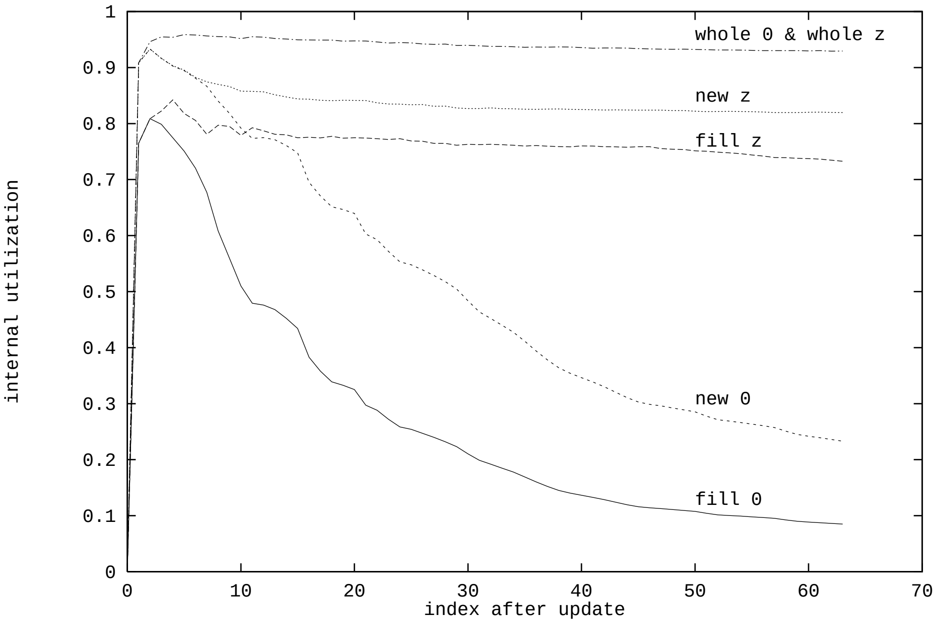

Figure 11: The long list disk utilization of the various policies. The x-axis is the database after the update. The y-axis is the fraction of space in the long lists which contain postings. Each curve is label with the values of *Style* and *Limit*.

the bottom two lines have *Limit* = 0 and the next two lines have *Limit* = *z* for the new and fill styles. This means that in-place updates *double* the number of I/O operations required. This is due to each in-place update needing a read and a write operation. The graph shows that the whole style requires more I/O operations than either the fill or new style, regardless of the use of in-place updates. Since the whole style costs one read and one write operation for each append of an in-memory list to a long list, whether an in-place update occurs or not, the whole style is the upper bound in number of I/O operations for any style. From the figure, we see that value for the final index for the whole style and for the fill and new styles with in-place updates are within 10% of each other. Thus, these policies use approximately the same time to build the final index.

Another measure of performance of a style is the long list utilization rate, namely the fraction of space allocated in long lists disk blocks that have postings. Thus we measure the *internal* utilization of the long lists.[2] Figure 11 shows the long list utilization rate for the index, measured at the end of each update, for the same set of policies as the previous figure. The spike for all curves between update 0 and 1 is due to the utilization rate of 0.0 when there are no long lists. We see that that utilization without in-place updates for the new and fill styles falls dramatically. This is due to the large amount of wasted space for small in-memory lists for these styles. Adding in-place updates to the new and fill style permits blocks to be more efficiently utilized as shown by the figure. The whole style has good utilization regardless of in-place updates since each list is stored contiguously.

Comparing the I/O performance of policies to the corresponding utilization rates, we see that

[2] The amount of *external* fragmentation is a consequence of the strategy taken for managing the free blocks on disk (cf. Section 3) which we do not study here.

Policy		Word Rank			
Style	*Limit*	10	100	1000	10000
fill	0	97	62	42	0
fill	z	88	38	4	0
new	0	62	60	41	0
new	z	60	57	8	0
whole	0 or z	1	1	1	0

Table 5: The read operations required to read long lists for words of a given rank in the frequency distribution for the final index. The word ranked 10th is "is" (134,132 postings), word 100 is "much" (57,964 postings), word 1,000 is "worse" (7,647 postings) and word 10,000 is "queens" (490 postings).

the two best performing policies in terms of I/O performance are unrealistic due to the resulting extremely poor utilization rates. Thus, the doubling of the I/O operations for update cannot realistically be avoided. In choosing among the remaining alternatives, if update performance is crucial then the new style with in-place updates is best, and if utilization is crucial than either of the whole policies is best.

Measuring query performance for a policy is difficult since the typical workload depends on the information retrieval model (IRM). For a typical boolean IRM, a query contains a few words (less than 50) and the words tend to be the less frequently appearing words since frequently appearing words do not discriminate strongly between documents. Thus we would expect many query words to reside in buckets for this model. For a typical vector space IRM, a query may be derived from a document, consequently the query contains many words (more than 100) and the words tend to be frequently appearing words.

To evaluate query performance for a boolean IRM, the number of read operations needed to read certain words (those with frequency rank 10, 100, 1000 and 10,000) in the final index are totaled by counting one read operation for each chunk. These counts are shown in Table 5. The table shows a huge range of values for words with rank 10 and 100. However, these words are unlikely to appear in a query since they do not discriminate among documents very strongly. For the word of rank 1000, the whole style has better query performance by a factor of 4 over the fill style with in-place updates and by a factor of 8 over new style with in-place updates. The fill and new style without in-place updates have relatively poor performance. Note that since the frequency distribution is inversely exponential, many words have long lists about the same length as the word ranked 1,000. The 10,000th ranked word is in a bucket and so it requires 0 long list read operations. During query processing for this word, a read operation is required to fetch the bucket for this word. Overall the whole style offers superior query performance.

Another method to measure query performance for a boolean IRM is given by DeFazio [2]. DeFazio ranks the words in the vocabulary in order of frequency and then divides the ranking into three frequency sections: high (90% of all postings), middle (the next 5%) and low (the remaining 5%). Each section is then weighed equally. A feature of this measurement is that it heavily weighs words with few postings.

For our experimental database, the top ranked 20,826 words are in the high frequency section,

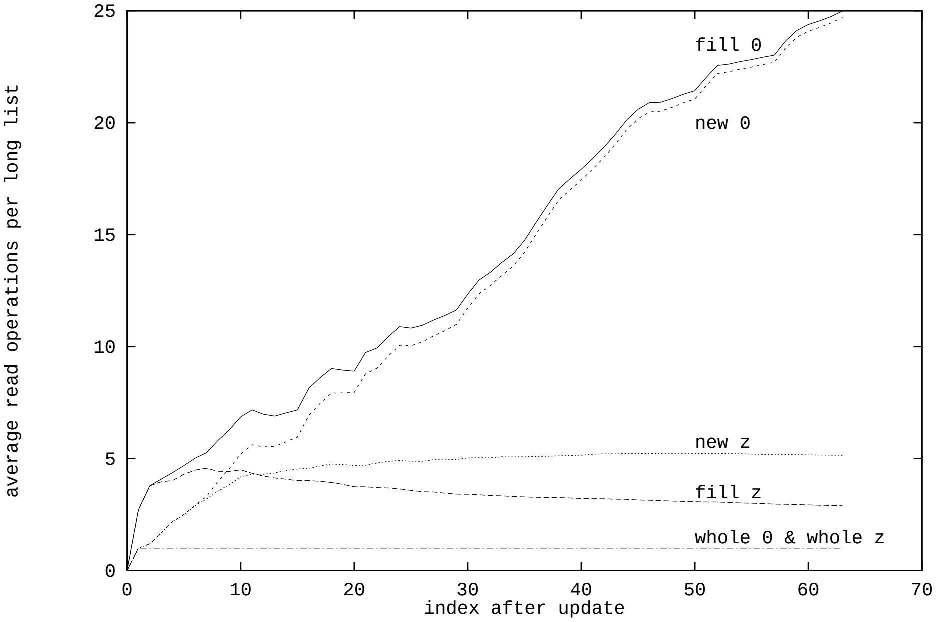

Figure 12: The average number of read operations to read a word with a long list. The x-axis is the database after the given update. The y-axis is the average number of read operations to read a long list.

the next 34,792 words are in the middle frequency section and the remaining 732,638 words are in the low frequency section. Figure 26 shows that in the final index there are 8,997 words with long lists. Assuming that all words with long lists fall in the high frequency section, then 43.2% (8997/20826) of the words in the high frequency section have a long list and the remaining words are in buckets. All of the middle and low frequency sections have words that are in buckets. Thus overall, 85.6% of the words for DeFazio's measurement resides in a bucket data structure since 2/3 of the time we choose words from the middle or low frequency sections. Thus, the long list allocation policies have little impact on the DeFazio metric.

For a vector space IRM, entire documents can be submitted as queries (e.g., for relevance feedback), so we expect the distribution of words in such a query to approximate the frequency of words in documents. To measure query performance for the vector space IRM, we measure at the end of each update the average number of read operations needed to read a long word. This is computed by counting the total number of chunks in the index and dividing by the number of words with long lists. Figure 12 graphs this result for the various policies. The graph shows that in-place updates are need for competitive query performance for the new and fill styles. In the final index, the whole style performs about 2.8 times better than the fill style with in-place updates and about 5 times better than the new style with in-place updates for this metric. Note that our qualitative conclusions from the table evaluation and the average read operations metric are the same, although they differ quantitatively. Clearly, if read performance is the crucial variable in the design of long lists, then the whole style is preferable.

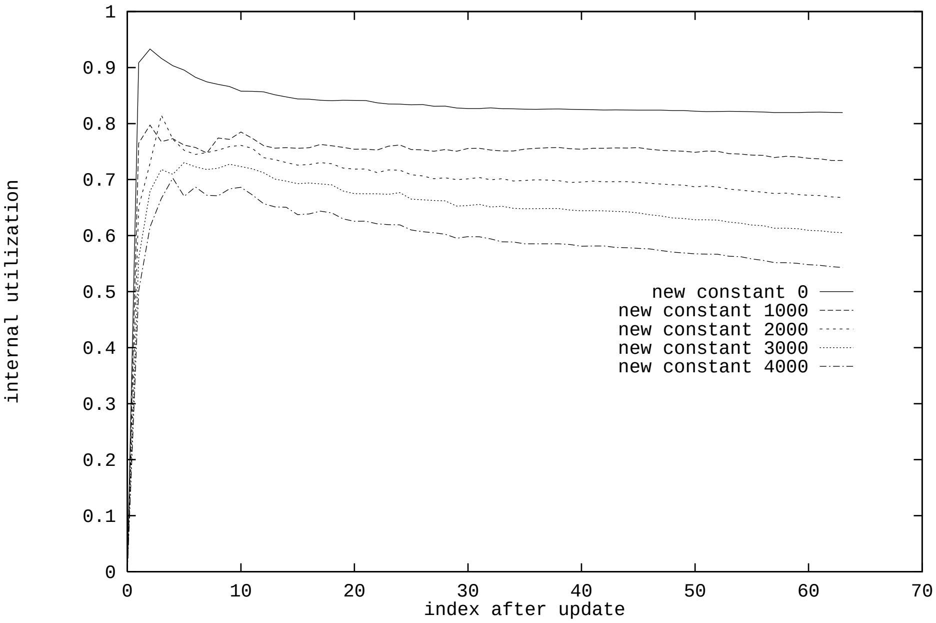

Figure 13: The long list disk utilization of the new style with a constant allocation strategy. The integer associated with each curve is the constant k in the allocation strategy.

5.2.2 Allocation Strategies

There are several issues to consider with the allocation strategies. How is the constant value for an allocation strategy selected? Given an allocation constant, is there some rule to select its constant, independent of a policy? Given any style, is one of the allocation strategies best? Let us start by focusing on a particular style, say the new style. (We assume in-place updates since allocation strategies are not otherwise used.)

For the new style, as the amount of reserved space for each list rises (by increasing k), the number of in-place updates rises and behavior converges towards a style where most updates long lists are in-place. In addition, as the amount of reserved space rises the disk utilization falls and the average number of reads for a long list approaches 1. This presents a classical trade-off between disk utilization and query performance. In-place updates also increase the update time for the new style, but the range of update times for in-place updates is only 10% in terms of I/O operations (cf. Figure 10). Thus, allocation strategies can only have a small impact on update time.

As an example of the space-time trade-off, consider the new style with the constant allocation strategy. As Figure 13 shows, as the amount of reserved space k is increased, the utilization rate of the long lists on disk drops indicating that the additional reserved space is partially empty. Simultaneously, the number of in-place updates that occurs increases, leading to a lower average number of read operations per long list, as shown in Figure 14. Note that there is a factor of 2 improvement in average read operations between a constant value of 0 and a constant value of 1000 with a corresponding approximately 10% loss in long list disk utilization. Other constant values offer much smaller improvements in query performance at a high cost of utilization. Thus, using a small value of 1000 achieves the highest benefit in query performance at the lowest cost of wasted disk space.

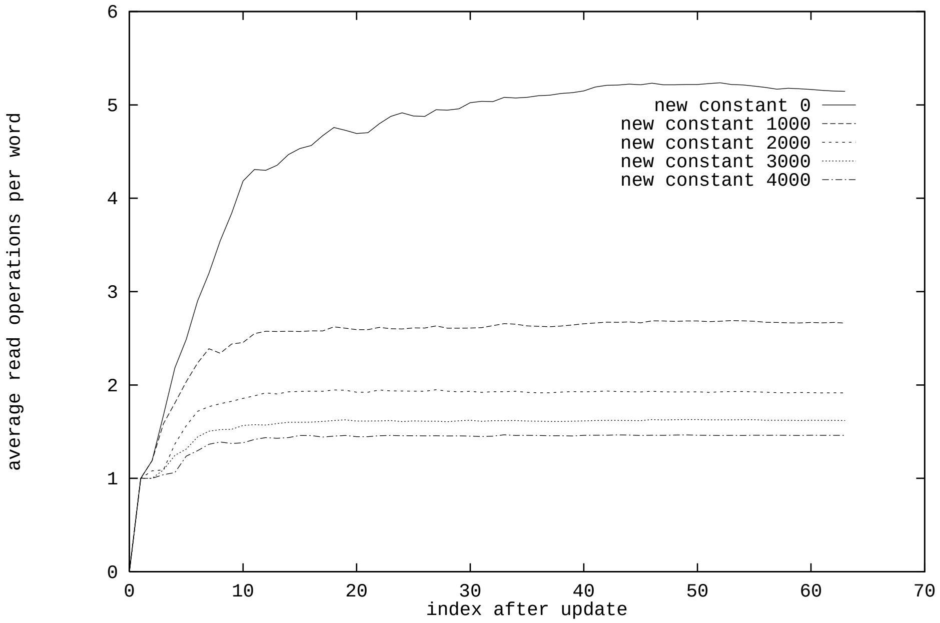

Figure 14: The average number of read operations to read a word with a long list for the new style with a constant allocation strategy. The integer associated with each curve is the constant k in the allocation strategy.

Given that all the allocation strategies have a space-time trade-off, which allocation strategy is best for the new style? Figure 13 shows that as the incremental updates progress, the utilization rate falls, where as Figure 14 shows that query performance stays the same. This is due to the decreasing size of new long lists which have overflowed from buckets in later incremental updates (cf. Figure 9). Since a fixed number of postings are reserved, for each overflow of a new long list with a smaller number of postings, the proportion of reserved postings grows, and thus the utilization rate falls. This property holds for any fixed size allocation strategy, in particular it holds for the constant allocation strategy (and the block allocation strategy which we consider below). However, the proportional allocation scheme does not operate this way, since as the size of the long list for new words decreases with each new word, the number of reserved postings decreases proportionally.

Table 6 compares various allocation strategies and constants for the new style. The "Read" column is the average number of read operations required to read a long list. The "Utilization" column is the internal utilization of the long lists. The "In-place" column is the total number of in-place updates needed to incrementally build the final index. The "Fraction" column is the fraction of in-place updates of the total possible number of in-place updates. (The total possible number of in-place updates which are possible is 213,256.) (The "In-place" and "Fraction" columns are included only for comparisons with Table 7.) The constant value for each strategy was chosen by increasing it until long list utilization was at 73%. This utilization rate was chosen since it offered good read performance, which was not available at higher long list utilization rates. Some additional values of interest are also included in the table. The table suggests (and other results not shown here confirm) that the new style with a proportional allocation strategy offers the best

New Style					
Allocation	*k* Value	Read	Utilization	In-place	Fraction
constant	500	3.60	0.78	189865	0.89
constant	1000	2.66	0.73	198292	0.93
block	2	3.36	0.78	192059	0.90
block	3	2.61	0.73	198746	0.93
proportional	2.0	2.89	0.80	196271	0.92
proportional	3.0	1.88	0.73	205316	0.96

Table 6: A comparison of allocation strategies with respect to the final index for the new styles. The meaning of each column is described in the text.

Whole Style				
Allocation	*k* Value	Utilization	In-place	Fraction
constant	0	0.93	179603	0.84
constant	500	0.90	188836	0.89
constant	1000	0.82	198309	0.93
block	2	0.87	194003	0.91
block	3	0.80	200156	0.94
block	5	0.68	205735	0.96
proportional	1.1	0.90	193695	0.91
proportional	1.25	0.85	202087	0.95
proportional	2.0	0.67	209994	0.98

Table 7: A comparison of allocation strategies with respect to the final index for the whole style. The table is describe in the body of the text.

trade-off by having the best read performance at this level of utilization.

Let us now turn our attention to the whole style. There is also a space-time trade-off for the allocation strategies for this style. The space trade-off is the utilization of long lists (as for the new style) but the time trade-off is only update time, not query performance, since all allocation strategies offer the same query performance. To compare update time, we cannot count I/O operations since this measure not distinguish between reading the tail of a list to append an in-memory list and reading the entire list. To account for this, we directly compare the number of in-place updates for each allocation strategy.

Table 7 shows statistics for various allocation strategies. The number of read operations for a long list is always 1.0 with the whole style. The "Utilization" column is the internal long list utilization. The "In-place" column is the total number of in-place updates needed to incrementally build the final index. The "Fraction" column is the fraction of in-place updates of the total possible number of in-place updates. (The total possible number of in-place updates which are possible is 213,256.) The table shows that the proportional allocation strategy is the best overall strategy since it is the only strategy to offer at least 90% for both utilization and the fraction of in-place updates.

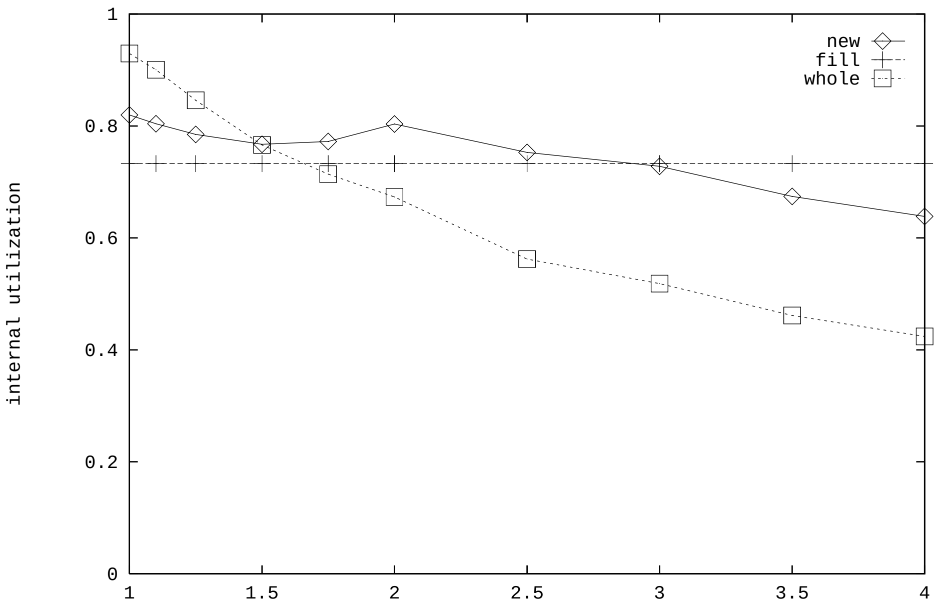

Figure 15: The impact of the constant k for the proportional allocation strategy on the utilization of long lists in the final index. The fill style (with the extent allocation strategy with extent size 3) is include for comparison.

Turning to the fill style, recall it has its own extent allocation strategy. The same space-time trade-off as with the new strategy exists. So as the number of extents e is increased, disk utilization falls and query performance improves. We conducted the same analysis for this style as for the others, increasing e until utilization falls to 73%. Our experiments show that a value of 3 for e gives an average number of read operations for a long list of 2.90, and 198,746 in-place updates at this utilization level. Both of these performance measures are worse than the best new style policy. However, the fill style as an advantage of limiting the maximum required contiguous region of disk (in this case to 3 blocks, since e has a value of 3).

We have seen that the proportional allocation strategy is a good choice for the new and whole styles. We now consider in greater depth the selection a good constant k for this allocation strategy. Figure 15 shows the impact of varying k on the utilization of long lists. The figure shows that, generally, as k rises, the utilization falls for both the new and whole styles (the fill style does not interact with the proportional allocation strategy and it is included for comparison). However, there is a cusp in the new style at a constant value of 2. This is due to the fact that multiple updates to the same word have approximately the same length. A constant value of 2 reserves space for one additional in-place update. The simultaneous increase in in-place updates is shown in Figure 16. We see that only a marginal improvement from 84% of the in-place updates at a constant value of 1.0 to 100% of the in-place updates at a constant value of 4.0 is possible. Consider both figures, we see the the majority of gains are from constant values less or equal to 3.0. In summary, based on the trade-offs presented, we choose the proportional allocation scheme a constant of 1.1 for the whole style and 3.0 for the new style.

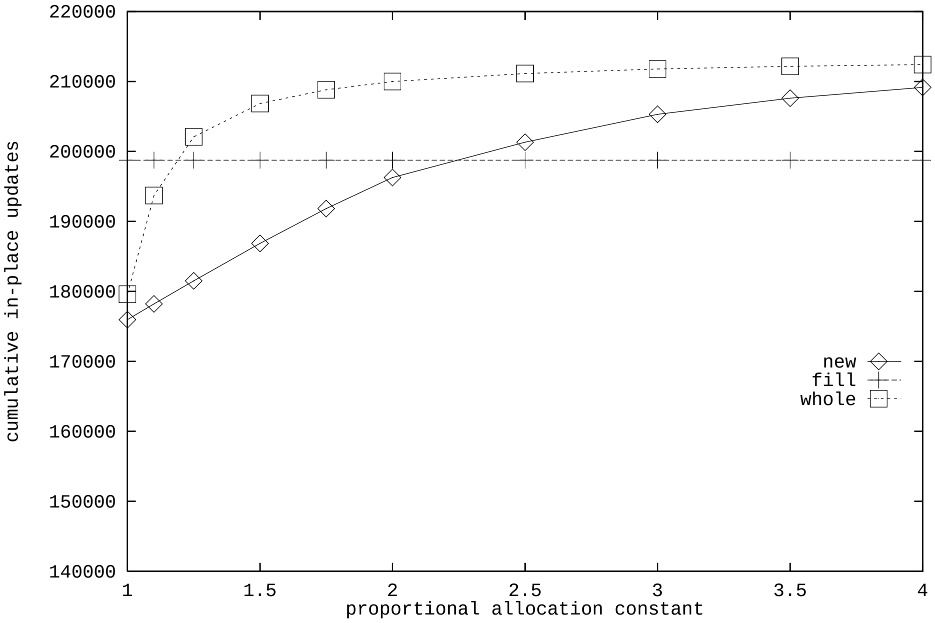

Figure 16: The impact of the constant k for the proportional allocation strategy on the cumulative number of in-place updates which occur in building the final index. The fill style (with the extent allocation strategy with extent size 3) is include for comparison. Note that the y-axis starts at 140,000.

5.3 Exercise Disks

In this section we compare the performance of the various allocation schemes by the actual times taken to execute a trace by the execute disks process. Note that the times in this section correspond to Phase II times as discussed in Section 4.6.

Figure 17 shows the cumulative time taken to incrementally build the final index. The fill style without in-place updates (fill 0) is not shown since our disks were not large enough to store the long lists for this policy due to gross underutilization of disk space (cf. Figure 11). Notice that the range of cumulative times for the final index vary by a factor of 4 as opposed to a factor of 2 determined by comparing total I/O operations (cf. Figure 10). The very significant difference between the different policies implies that a policy must be chosen with care.

Comparing Figure 17 with Figure 10, we see that measuring cumulative I/O operations produces the same *qualitative* comparison of policies as measuring real execution time. That is, the ordering of policies from best to worst is the same (accounting for the addition of whole with in-place updates and the removal of fill without in-place updates). This confirms our use of I/O operations to compare policies.

We also see that the new style with limit of zero policy has an almost linear growth in the cumulative time taken as opposed to a more steep increase in the cumulative number of I/O operations. This is due to the coalescing of I/O operations by the exercise disk process. That is, since for long list updates this policy only writes sequentially to the disk, all the write operations in an update can be coalesced (up to the buffer size imposed by the exercise disk process). Figure 17 also shows that the whole style without in-place updates takes the longest cumulative time to build

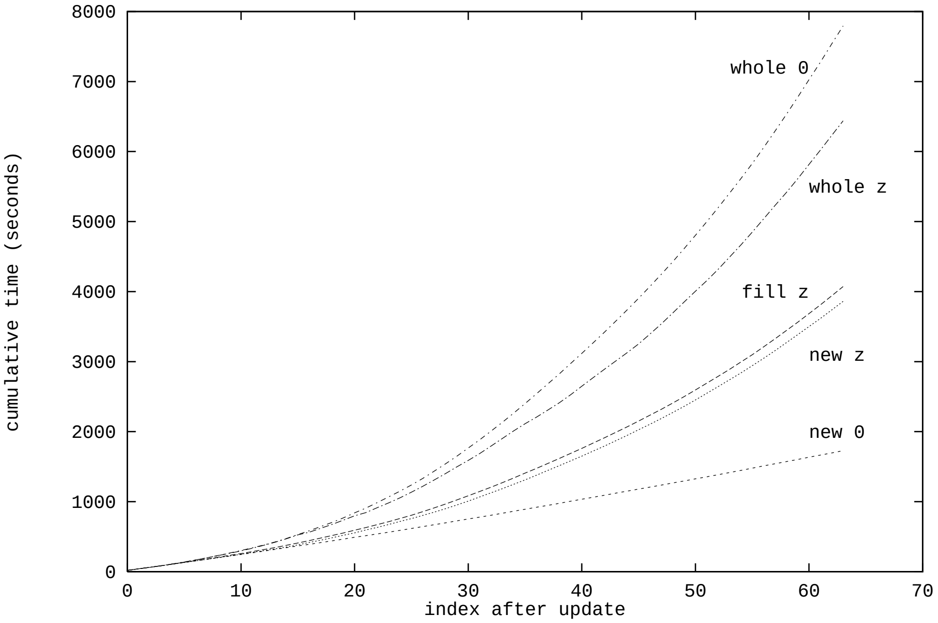

Figure 17: The *cumulative* time needed to build the final index. The x-axis is the index after the given update. The y-axis is the cumulative time needed to build the index incrementally. Each curve is labeled with the values of *Style* and *Limit*.

the final index. This is due to the additional movement of long lists compared to in-place updates.

Figure 18 shows the time taken to perform each update. That is, this figure is the non-cumulative version of the previous figure. The update times grow over time as the number of long lists in the index grows. However, the increase for new style without in-place updates is slight because updates to different long lists are coalesced into single I/O operations. A second effect shown in the figure is that the whole style with in-place updates ($Limit = z$) is the only policy whose per update time is sensitive to the variations in the size of the update. The average number of in-place updates for this policy is sensitive to the average number of postings in a long list update.

5.4 Policies

This section has compared the performance of the various options for storing long lists. Generally, the schemes that rewrite the unused space at the end of a long list before allocating more space take considerably longer than the schemes that don't rewrite the space, due to the extra read required for each long list. However, the extra space consumed by *not* rewriting the tail of long lists makes that option impractical for most applications.

While we have analyzed various situation, a designer of an IR system is faced with the issue of chosing a policy. Is there a best policy? In general, no policy is best. Some policies favor update time and others favor query time. We consider each policy in turn.

new style: The new style provides the best update performance, since only the last block of each long list need ever be read during the updates. The new style without in-place updates is the

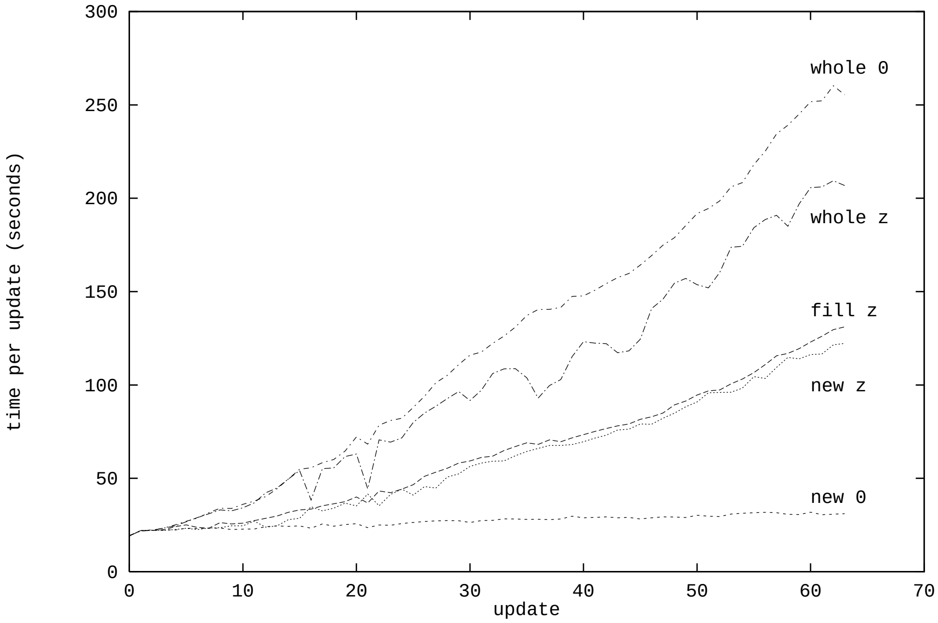

Figure 18: The time per update. The x-axis is the update number. The y-axis is the time to execute the update by the exercise disk process. Each curve is labeled with values of *Style* and *Limit*.

best if update time is critical. The policy offers a factor of 2 in update time over the next best policy and a factor of 4 over the slowest policy. However, the policy exhibits very poor query time and disk utilization. The new style with proportional allocation with constant 3 compensates for the poor query time and disk utilization. This policy is best if update time is important with reasonable query time. The policy is faster by a factor of 2 over the slowest policy and offers query performance within a factor of 2 of the best query performance.

Bottom Line: Use the new style only if query performance is not critical. If you do use the new style, we recommend a proportional allocation strategy with a constant of 3.0.

fill style: Essentially the fill style offers no advantages over the new style except that the maximum contiguous section of disk required is limited (it is unbounded in the new style). This requirement has an advantage in that long lists are automatically divided into sections of disks which can be written and read in parallel (e.g., with a disk array). The cost of satisfying this requirement is small in the case of the fill style with extent allocation with constant 3 since this policy has slightly worse performance on all three metrics than the new style with proportional allocation with constant 3.

Bottom Line: Performance is comparable to the new style; better for disk arrays. If you do use the fill style, we recommend 3 extent blocks.

whole style: The major advantage of the whole style is its guarantee of 1 read operation for any long list. Providing this guarantee has a cost in index build time. The penalty arises from the costs of moving long lists to keep them sequential. However, this penalty is not as large as might be expected, due the relative efficiency of performing sequential disk reads and writes.

Variable	Value	Description
$Seek$	20.0	Seek time (ms)
$ReadBlock$	3.8	Read a block (ms)
$WriteBlock$	4.0	Write a block (ms)

Table 8: The variables controlling the simulation of the exercise disk process.

The whole style without in-place updates has an about 12% slower update time compared to the whole style with a proportional allocation with constant 1.1. The latter policy has a long list utilization rate of 90%. Thus the latter policy is the best if query time is critical.

Bottom Line: Use the whole style if query performance is critical. For this style we recommand a proportional allocation strategy with a constant value of 1.1.

6 Extensions

In this section we describe some extensions to the experimental design described in Section 4. The purpose of the extensions is to predict the behavior of the dual-structure index in a wider class of scenarios, e.g., as the database grows beyond the size of the experimental text document database or as more disks are added.

To accomplish this, we replaced the database and the invert index process with a simulator process which generates in-memory inverted index batch updates directly. In addition, we replace the exercise disk process with a *synthetic disks* process which uses synthetic functions to estimate the I/O times of a collection of read and write operations. Replacing the exercise disk process allows rapid exploration of the parameter space and an expansion of the parameter space to include disk performance parameters and a variable number of disks.

6.1 Synthetic Disks

Using real disk to execute traces gives exact results, but a run takes from a half hour to over two hours. Figure 17 takes about 6 hours and 40 minutes to generate! Synthetic disks are much faster to process and can offer almost as good accuracy if they are tuned properly.

Our synthetic disks are simple analytic functions. Our model is based on an average time to read or write the first block of a request, plus an additional charge for each subsequent block in the request. This model takes into account the relatively large time to seek to a block and the relative efficiency of reading or writing many blocks at a time.

Table 8 lists the variables used in the synthetic functions. Let x be the number of blocks to read or write. For read operations, we use $Seek + ReadBlock \cdot x$ as the number of milliseconds required for the read operation. For write operations, we use $Seek + WriteBlock \cdot x$, as the number of milliseconds required for the write operation. These parameters were derived from measurements of three disks on a single controller for workloads like those used in this study. The synthetic disk process reads a disk trace, computes the time taken by each disk used the read and write formula, and then reports the maximum of these times as a prediction of the total time need to execute the disk trace by the exercise disk process.

To verify the accuracy of the model, we compared the actual times encountered in running the

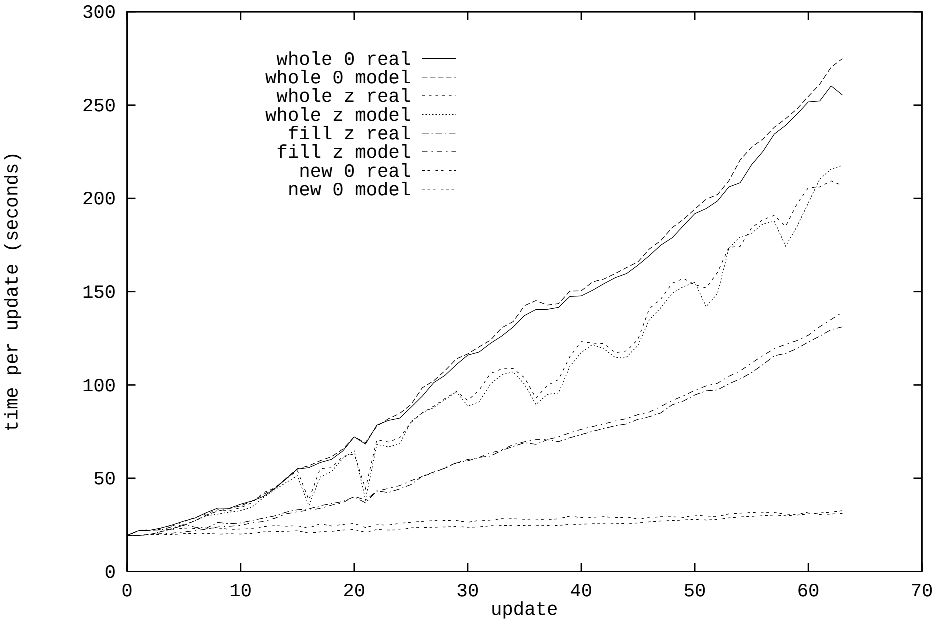

Figure 19: Comparison of the exercise disk process times to the synthetic disk model times for each batch update. Each line labeled real is the exercise disk process. Each line labeled model is the synthetic disk model.

disk exerciser on several styles without allocation policies (we used the constant allocation policy with 0 reserved postings) to those predicted by the disk model. The results are shown in Figure 19. We see from the graph that the synthetic disk process does a good job of estimating the time taken by the exercise disk process.

To compare seek times, we modeled disks with 10, 20, and 50 millisecond times to read or write the first block plus the data rates that we measured on our real disks. The results of this comparison are shown in Figure 20. There is a strong component of seek time even in the whole scheme. Initially, most of the I/O operations are for rewriting the buckets and are strongly sequential. As the trace progresses, the average size of the data of an I/O operation falls to only a few blocks as more long words are created. Since the average number of blocks written for each long word is small, the average size of the data of an I/O operation falls and the seek time begins to dominate the total I/O time required to add a batch to the data base. Since the number of seeks is proportional to the number of long words, the faster seek time disks have a smaller slope of increased update time as batches are added.

To compare the effects of data rate, we model disks with 20 millisecond seeks as seen on our real disks and used a variety of data rates, as show in Figure 21. The slow transfer curve has a data rate based on optical disks (660 KB/sec reads, 210 KB/sec writes). The medium transfer curve has a data rate based on our disks (1 MB/sec to read and write). The fast transfer curve has a data rate based on current high end disks (8 MB/sec to read and write). Two effects are visible from this figure. First, the slower data rate device exhibits far larger variations in update times. The spikes in the update times are due to update batches that move many long lists. Faster transfer times also significantly reduce the update times. This variation is mostly due to the reduced time

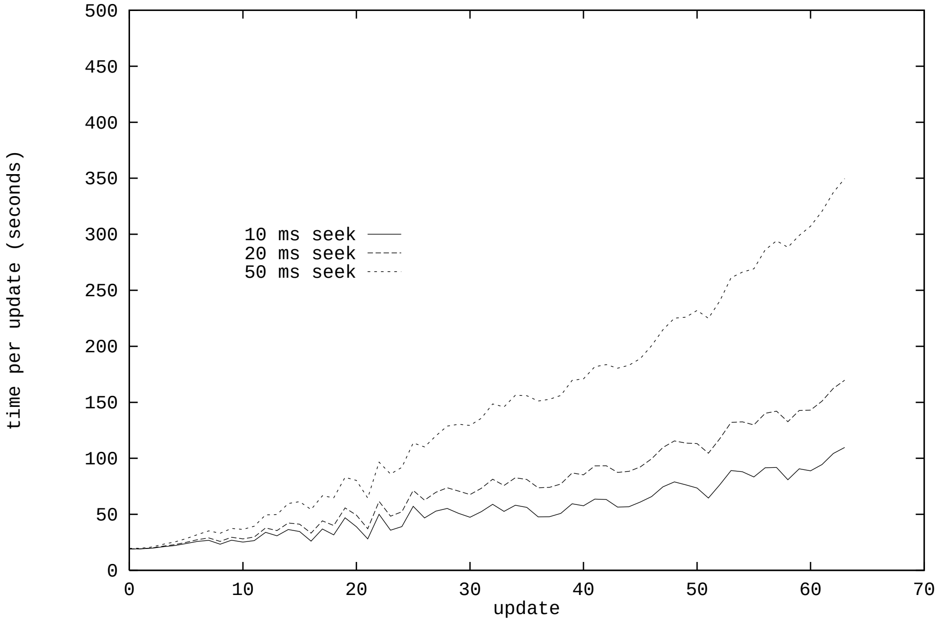

Figure 20: Comparison of varying seek times. The whole style with proportional allocation scheme is the policy used.

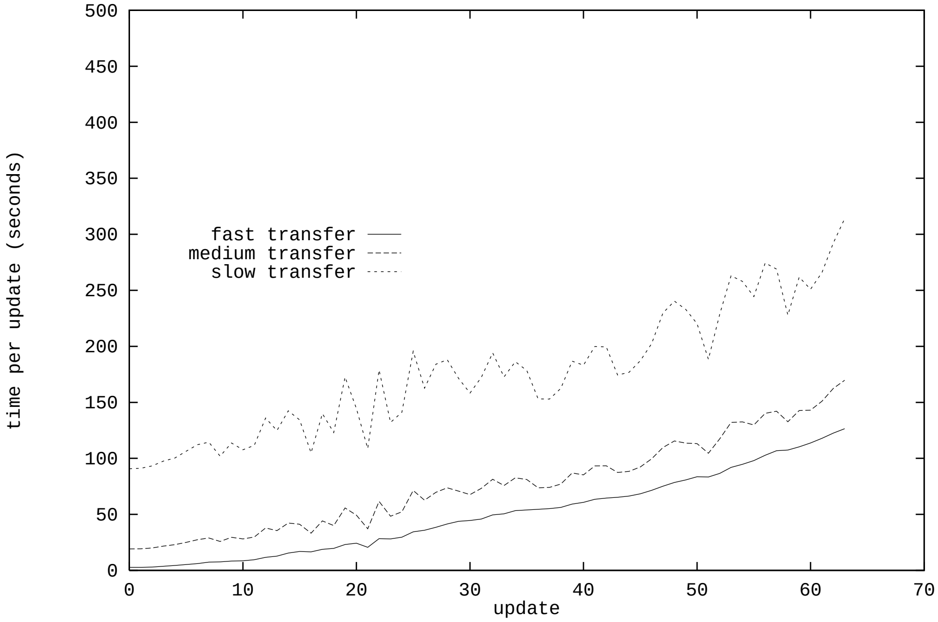

Figure 21: Comparison of varying data rates. The whole style with proportional allocation scheme is the policy used.

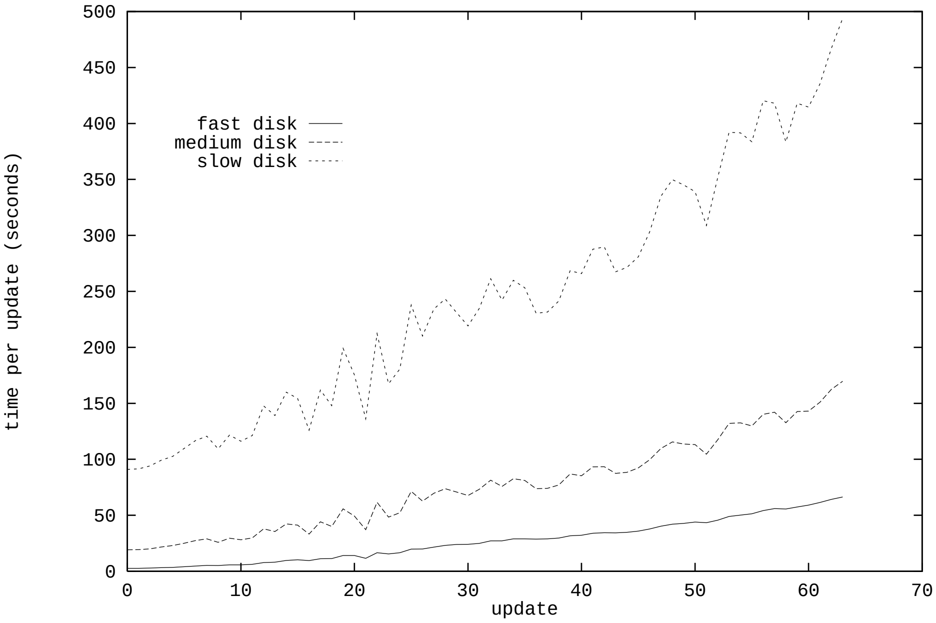

Figure 22: Comparison of varying disks

to rewrite the buckets on each disk. That difference is seen on the time to required to process the first batch.

To provide a complete picture of the differences in index time caused by disks of various performance, we compare the various devices directly. The slow disk curve represents optical drive parameters (50 millisecond seek, 660 KB/sec read, 210 KB/sec write). The medium disk curve represents our disk parameters (20 millisecond seek, 1 MB/sec read and write). The fast disk curve represents the latest generation disk parameters (10 millisecond seek, 8 MB/sec read and write). The results are shown in Figure 22. Naturally, the effects of the previous two studies are combined. The fastest seek time provides a more gradual slope coupled with the fastest data rate giving the lowest base update time.

Another important variable in the configuration of a system is the total number of disks. To study the effect of this variable we fix the single disk parameters to those of our current disk drives. Note that these parameters incorporate our current configuration of 3 drives on one controller. Thus, in this study, all the drives are assumed to be grouped 3 per controller. Once again, the whole style proportional allocation strategy with a constant 1.1 is used and the simulation was driven by our real text document database. Given a properly spread I/O load, the use of multiple disks simulates a smaller number of disks that support a high data rate and fast seek time. This effect is borne out in this study, shown in Figure 23. As more disks are used, the slope of the lines flatten out and the variations in batch time are reduced, as seen in the seek study. In addition, more disks reduce the base time to rewrite the buckets, as seen in the data rate study.

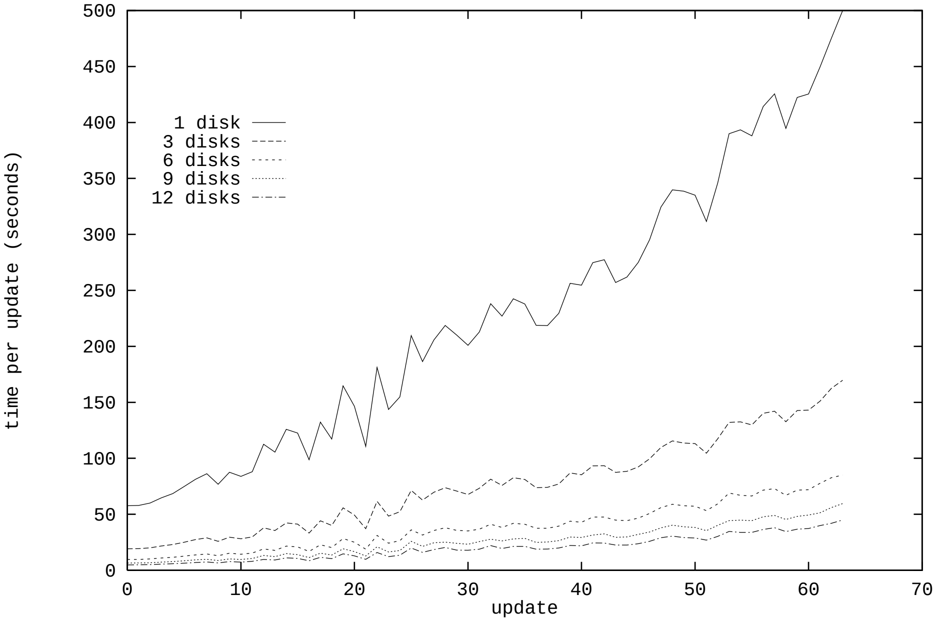

Figure 23: Comparison of varying numbers of disks for the whole style. Each set of three drives has an I/O controller.

6.2 Batch update generator

One of the strengths of our study is the use of real documents. However, our database represents only 64 days of data occupying 646 MB. While this volume is sufficient to study many issues, some behavior of our algorithms appears only with larger databases. We generate larger databases by synthetically generating batch updates. We accomplish with the following steps.

1. Estimate the growth in the vocabulary of the text document database.
2. Construct a word-occurrence distribution for the synthetic final index.
3. Estimate the number of documents in a batch update.
4. Estimate the number of unique words in a document.
5. Synthetically generate 200 batch updates.
6. Verify that the generator is accurate.

To estimate the growth in the vocabulary of the database, we note that the number of new words that are introduced in a batch does not go to zero over time. This is because new words are constantly introduced to the database–family names, company names, misspellings, etc. Thus, in each batch we generate words that are already in the index and synthetic words that are new to the index.

In Figure 24, the data points graphs the size of the vocabulary measured at the end of each batch update. The x-axis is the number of postings and ends at about 48 million postings, which is the total number of postings in our database. We use postings as the axis instead of the batch

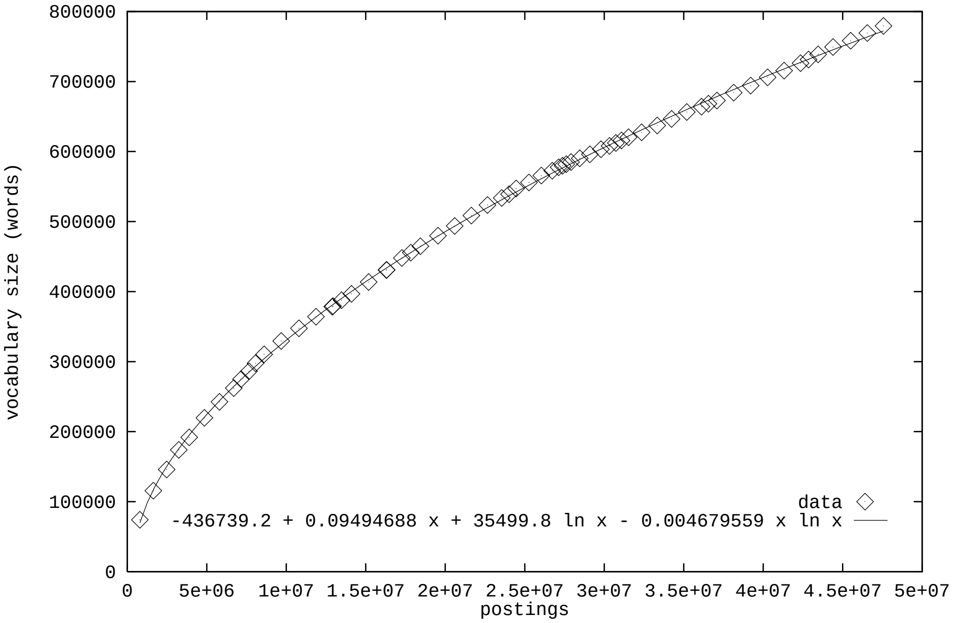

Figure 24: Comparison of the synthetic vocabulary size function f to the actual growth in vocabulary size. Each data point is generated by counting the vocabulary size of the index after each incremental update.

update number because we expect words to be introduced into the vocabulary on a per posting basis instead of on an arbitrary division of documents into batchs. The y-axis in the figure is the number of words in the vocabulary. Each data point is the size of the vocabulary after the given incremental update. The figure shows the continued introduction of new words into the index for each incremental update.

To predict the size of the vocabulary as the number of postings grows beyond 48 million postings, we generate a function $f(x)$ by a curve fit to the data. To determine the form of f, we start with the function $x \ln x$ (which is analytically derived in [1]) and then add all combinations of lower order terms. The result of fitting the data to the equation $a + bx + c \ln x + dx \ln x$ using [10] is the equation

$$f(x) = -436739.2 + 0.09494688x + 35499.8 \ln x - 0.004679559x \ln x.$$

(We also tried curve fits for every subexpression of the function. The function above had the lowest least-squares error of all the curve fit equations.) Figure 24 shows a close agreement between the growth in the vocabulary size and the function f. However, the curve fit equation has a limited range of utility since it eventually produces negative numbers as x approaches infinity (due to a negative value of d – the constant for the dominate term of the function). We limit the scaling of our database to just over 3 times the number of postings of the final index. Within this range the curve fit function behaves reasonably.

Next, we construct the word-occurrence distribution for a *synthetic final index* consisting of *TotalPosting* postings based on

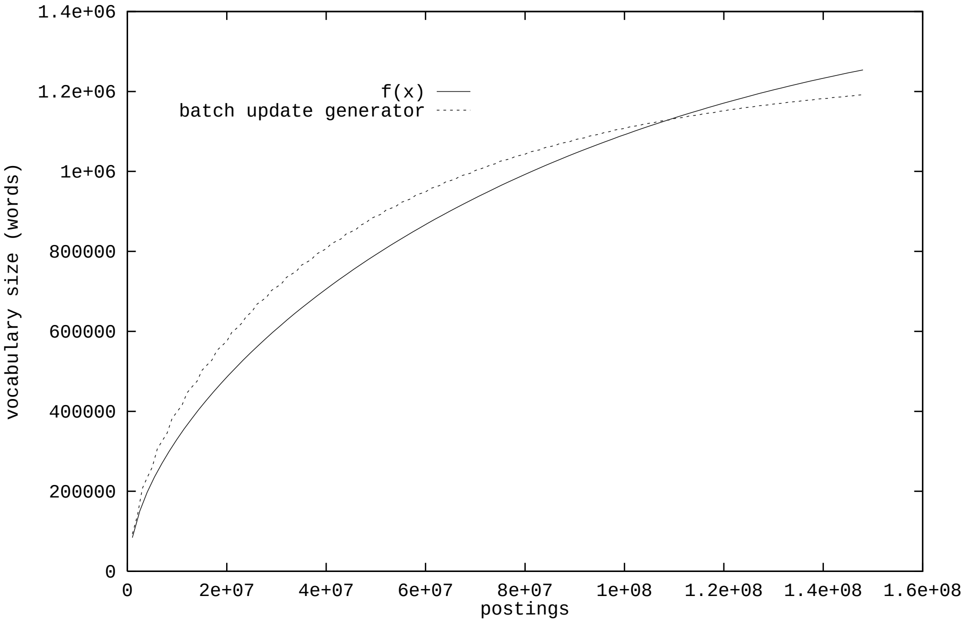

Figure 25: Comparison of the batch update generator to the curve fit function.

- f to estimate the number of unique words in the synthetic final index,
- the distribution of the final index in the experimental database, and
- assigning 1 posting occurrence to each new word synthetically generated (i.e., a word that does not appear in the experimental database).

We will use this distribution to model the probability that a word appears in a document. To compute *TotalPosting* for our synthetic final index, we linearly scale the experimental final index of about 48 million postings for 64 batch updates to predict *TotalPosting*= 151 million postings for 200 updates. Using f, the synthetic final index has an estimated total vocabulary of about 1.26 million words.

To generate a batch, we assume that the number of documents per batch and the number of unique words per batch are about equal to the averages of the 64 known batches, so we let each batch have 2,124 documents and each document have 350 unique words. Then, random documents are created by selecting random words from the synthetic final index until a desired number of unique words have been selected. The probability that a word is chosen is proportional to its frequency distribution, e.g., if the word cat occurs x times, then its probably of being chosen is $x/$ *TotalPosting*. From the random documents a synthetic batch update is generated.

To verify that the batch generator is accurate, we compare the vocabulary size of the function f with the vocabulary size produced by the generated documents. Figure 25 shows that the batch update generator initially overestimates the vocabulary size. This is due to the fact that the selection of a word for a batch update is over the entire range of words in the final synthetic index. Eventually the generator underestimates the vocabulary size. This indicates that not every possible word was chosen for some batch update. However, we believe the generator is good enough to qualitatively compare the behavior of databases scaled to larger sizes.

Buckets	*BucketSize*	Fraction
30	325000	0.798535
150	65000	0.799382
750	13000	0.804081
1500	6500	0.809564
7500	1300	0.843670
15000	650	0.863982
75000	130	0.910547

Table 9: The trade-off between the number of buckets and the size of the long lists for a fixed size bucket data structure. The "Fraction" column lists the percentage of words and postings that are in long lists for the final index.

6.3 Tuning the bucket data structure

We return for a moment to the experimental text document database as our basis for studying the behavior of the bucket data structure. Choosing a bucket size and a number of buckets is tricky. The best size of the bucket data structure depends on the amount of data indexed from the original database. Stop lists, stemming algorithms, and the granularity of the index (e.g., document level, paragraph level, sentence level, or word level.) all strongly influence the amount of data indexed. Choosing a bucket data structure that is too small will lead to poor performance due an inordinate number of overflows. Choosing a bucket data structure that is too large will lead to inefficient utilization of memory and poor performance due to the cost of writing the bucket data structure to disk. We do not consider these issues in detail, but we do make a few remarks. Note that whatever organization for a bucket we choose, this decision does *not* impact the comparison of long lists policies. This is because the bucket data structure *uniformly* affects the long list policies.

For a fixed size bucket data structure, as the number of buckets increases the size of a bucket decreases. Smaller buckets overflow words more quickly than larger buckets, leading to an overall increase in the number of long words. However, small buckets have the advantage that they can be read more quickly during query processing. As the number of buckets increases, the number of long lists in the index increases, so the time to update also increases. Table 9 shows the increase of long words and postings in the final index as the number of buckets increases. We use a much smaller value (1,500) to keep the number of long lists low. An even smaller value would create very large buckets which would negatively impact query processing performance.

So far we have assumed a fixed bucket space. Our dual structure algorithm works well in that it readily adapts to the available space. However, after many updates, shorter and shorter lists become long lists. This suggests that the bucket data structure requires *expansion*, i.e., an increase in the memory committed to the bucket data structure. This raises a host of issues – should there be more buckets or larger buckets? periodic or continuous expansion? a threshold value to trigger expansion? based on the size of the list or the fraction of the index in buckets, or some other criteria? However, whatever expansion scheme is chosen, it is not critical for comparing allocations schemes or the other issues presented here.

We consider one possible expansion here. We limit the number of long lists by incrementally increasing the amount of memory dedicated to buckets as the index grows. Figure 26 shows the

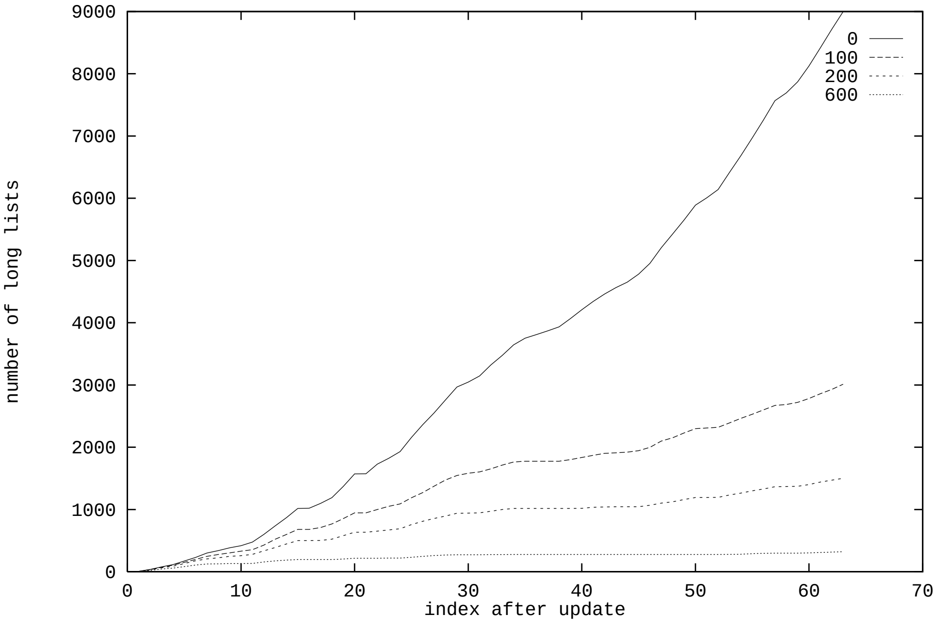

Figure 26: Controlling the creation of long lists by increasing the bucket size. The x-axis is the update number. The y-axis is the number of long lists in the index at the end of the update. As updates are processed, the bucket size is determined by the function $BucketSize + dx$ where d is the label associated with each curve and x is the update number.

number of words with long lists at the end of each update. Each curve corresponds to the function $BucketSize + dx$ (where d is the *delta* associated with a curve). For instance, the second to top curve is labeled 100, so the size of the buckets at each update x is $6500 + 100x$. The graph demonstrates that the number of words with long lists in the index can be held essentially constant as the index grows by allocating a constant increase in the size of the bucket data structure.

The memory cost of increasing the memory dedicated to the bucket data structure is high. With a delta of 0, the total size of the bucket data structure is $Buckets{\cdot}(BucketSize{+}dx) = BucketTotal$ or $1500 \cdot (6500 + 0 \cdot 63) = 9,750,000$ words and postings. At the end of the last update, the bucket data structure holds 19.8% of all the words and postings. Assuming that a word or a posting needs 6 bytes to store on average, the bucket data structure is 10.9% of the size of the raw text of the database. With a delta of 100, the total size of the bucket data structure is $1500 \cdot (6500 + 100 \cdot 63) = 19,200,000$ or almost double the size of the bucket data structure with a delta of 0. We use a delta of 0 for the remainder of the paper.

6.4 Scaling Bucket Behavior

We now apply our synthetic batch generator to examine the behavior of buckets as the index is scaled in size. Consider the whole style with a proportional allocation scheme with constant 1.1. We use the batch update generator and the synthetic disk process to estimate the per batch update time for 200 batch updates. Figure 27 shows this situation for several different sizes of the bucket data structure. (We vary the number of buckets and keep the size of each bucket constant.) Examining the curve for the first 20 updates, we see that as the bucket data structure grows in

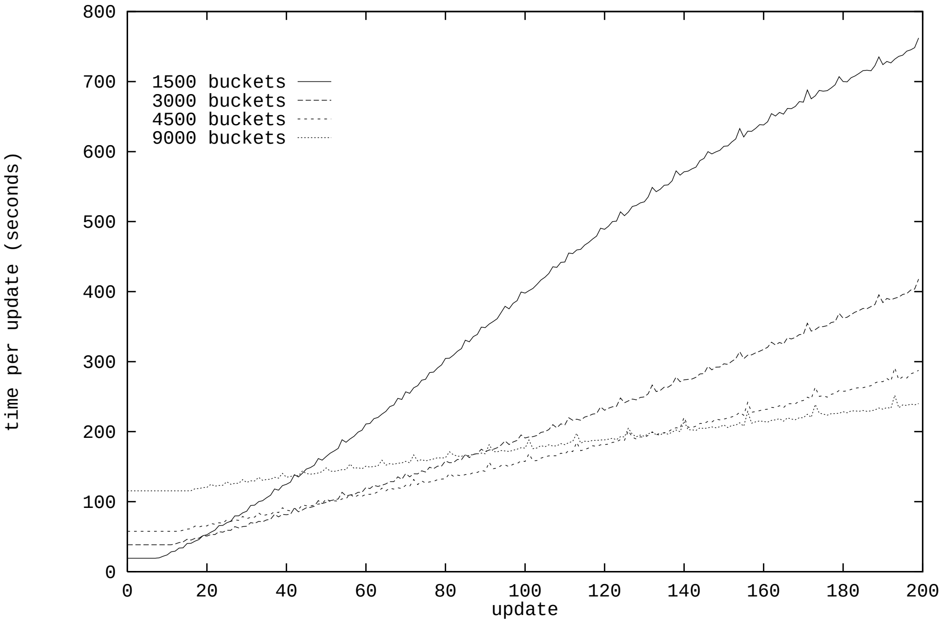

Figure 27: The update time per batch for the whole style with proportional allocation scheme for various bucket sizes.

size, the update time per batch is higher. This is due to the need to write out a larger bucket data structure. Over the entire set of updates, however, the larger bucket data structure eventually has the lowest per batch update time since the smallest number of long lists exist for the index with the largest number of buckets. The spikes in the curves correspond to an unusual amount of moving of lists for that update. This is to be expected since each batch update is of a fixed size. This means that many of the long lists are growing at a constant rate and consequently they all overflow the 10% reserve space at the same time. The slight bend in the curve for 1500 buckets is due to the flattening out of the vocabulary size for the index as it grows.

For the same situation as the previous figure, Figure 28 shows the space occupied by the index for each of the bucket sizes. The jump in the space occupied in each curve is due to the update in which the largest lists overflow from the bucket data structure. Since each of these very long lists has a 10% reserved space, a tremendous amount of space is reserved for these lists. We see that the slope of the curve for the 1500 bucket scenario is higher than the slope for the 9000 bucket scenario. Eventually, these curves will converge to points separated only by the size of the bucket data structure. This is due to the bucket data structure filling up with words containing short lists with only a few postings.

6.5 Full text indexing vs. abstracts indexing

Some information systems index all occurrences of a word, rather than just the count of the occurrences in each document. Indexing all occurrences supports efficient query proximity operators such as finding terms adjacent to one another or within some specified number of words of each other. To determine how effective our algorithms are for this class of systems, we model the index for a full text indexing system. First, we assume that the position information for each posting

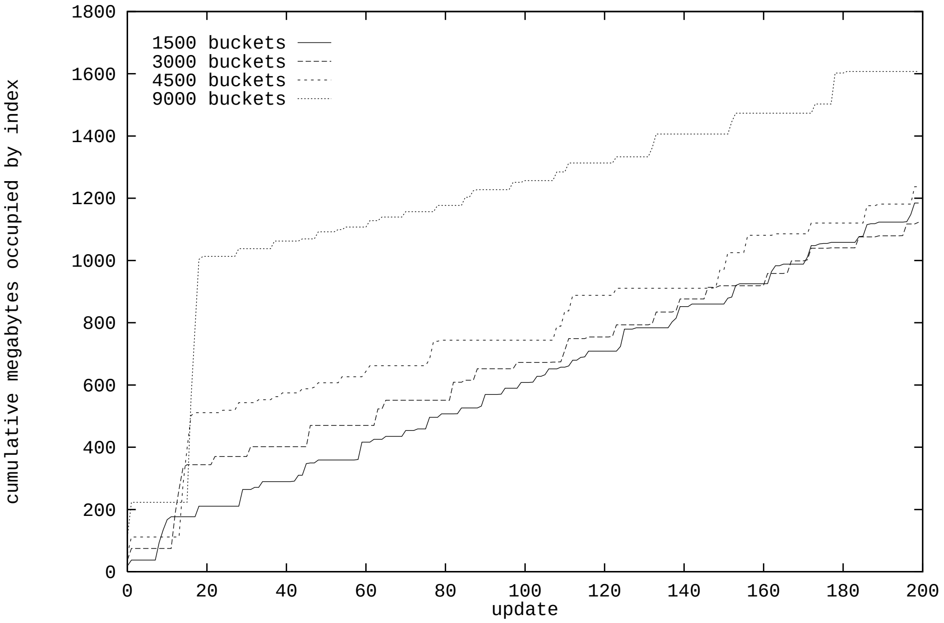

Figure 28: The update time per batch for the whole style with proportional allocation scheme for various bucket sizes.

would take the same amount of space as the count information, so the size of individual postings is unchanged. Second, we model the increase in the number of postings by adding code to the batch update generator to index every occurrence of a word. Third, actual systems often limit the increase in postings for this kind of index by using a "stop list" of very frequent terms that are presumed to offer limited search discrimination. To model the effects of a stop list, we eliminate the top 20 most frequent words[3] from our synthetic trace. The following table summarizes the posting volumes seen when indexing one occurrence per document, all occurrences, and all occurrences with a stop list for our synthetic trace of 200 batches:

Type of index	Posting count	Index space	Total build time
One occurrence per document	148 million	1.1 GB	7.0 hours
All occurrences	216 million	1.7 GB	7.9 hours
All occurrences with stop list	163 million	1.4 GB	7.2 hours

As can be seen from the table, indexing all occurrences adds 46% more postings. The use of a stop list reduces the postings about 25%. (Using a stop list on the one occurrence per document index only reduces postings about 5%.) In addition, indexing all occurrences increases the variation in long list sizes.

Figure 29 shows the time taken to index the three alternatives. These results were produced with the whole proportional style with a proportion of 1.1. A total of 4500 buckets instead of the default 1500 buckets is used since the trace is for three times as many batches as for the real document data. We modeled the disk behavior for three disks with the performance of our real

[3] The words: newsgroups, from, subject, a, in, of, to, the, and, is, that, for, it, on, s, with, be, i, this, and have.

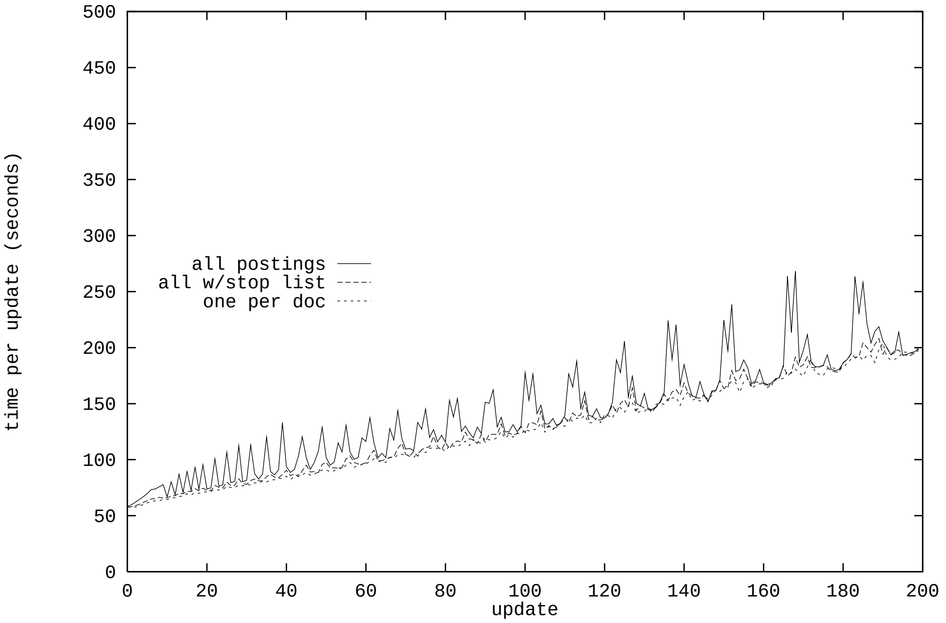

Figure 29: Times per update on synthetic documents comparing indexing one occurrence per term per document, all occurrences of a term, and all occurrences of a term minus a 20 word stop list.

disks. The large variations in the time per update are caused when many long lists are moved on the same update. That is, since the batch update generator produces updates of the same size, may lists overflow the 10% reserve space at the same time, causing a move of the long list instead of an in-place update, which leads to spikes in the time per update.

Note that the use of the 20-word stop list removes the wide variations in batch build time. There is still a penalty of 27% additional disk space consumed, but a total build time only 3% larger. The reason that the build time does not increase that much is that the time is dominated by the number of seeks. Since the number of long words updated on each batch is about the same, there is little variation.

In summary, we considered extensions to the basic study to see how our index designs behave with considerably larger text document databases, with various disk performance, with various numbers of disks, and with full-text indexing versus abstracts indexing. We found that the algorithms scaled with larger databases, as long as sufficient bucket space is allocated. However, additional bucket space adds a fixed cost to each update, so larger databases do lead to a modest increase in build time.

For disk performance, we found that faster disk data rates reduce the base cost per update due to rewriting the buckets. Faster data rates also reduce the impact of moving long lists in the *whole* scheme. Faster seek times reduce the slope of the curve as the text index grows due to an increasing number of long words. We also found that our algorithms use multiple disks well, with significant reductions in build times as more disks are added to the system.

For full-text indexing, we found a significant variation in batch build times as the top few long words overflowed their buckets. The use of a stop list to eliminate indexing of the most popular words removes these large variations in build times. With the use of a stop list, only a slight increase in build time is observed relative to abstracts indexing, although a significant increase in disk space occupied by the index is observed.

7 Related Work

Cutting and Pedersen [1] consider incremental updates of inverted lists where a B-tree is used to organize the vocabulary. Updates are optimized by storing short inverted lists directly in the B-tree. In our framework this optimization can be represented by a very small bucket for approximately each word in the text document database. However, in Section 5 we show that using few, larger, buckets offers better performance. In addition, our scheme dynamically determines a threshold value for determining if an inverted list is stored in a fixed sized structure or a variable length one. Cutting and Pedersen also described a buddy system for the allocation of long lists. This approach deserves further experimental study since it offers comparable space utilization and it is not clear if it offers better update performance than the methods presented here.

Faloutsos and Jagadish [4] extensively analyze the physical organization of long list. They study three methods that correspond to our whole style with a proportional allocation scheme, our new style with an adaptive allocation scheme (not studied here), and an unique style that combines benefits of the whole and new styles. Performance comparisons between our work and the schemes presented there are difficult since updates are not batched in that paper.

In another work, Faloutsos and Jagadish [3] extensively analyze a dual-structure scheme based on signature schemes for long lists and inverted lists for short lists. The division in the structure is static as opposed to a dynamic scheme presented here. In addition, the we believe that using inverted lists for short lists is computationally expensive since many I/O operations, each containing only a few postings, are required to update this structure.

Zobel, Moffat and Sacks-Davis [12] consider several issues in inverted file indexing. The compression methods presented there complement this paper well. They also consider fixed size buckets for storing inverted lists but do not discuss techniques for handling long lists. To compare approaches, a linear scaling of the 45 minutes update time of 132 MB of documents cited in the paper to the 686 MB text document database used here gives an update time of about 233 minutes. We halve this number to 116 minutes to account for improvements in processor performance. Our study predicts a range of index build times from about 43 minutes to 173 minutes depending on the policy used.

An interesting and entirely different approach, by Fox and Lee, based on preprocessing of document representations and a merge update of inverted lists is described in [5]. A non-incremental update time of 1 minute 14 seconds for 8.68 MB of documents appears in this article. Harman and Candela [6] also describe an update method and cite an indexing time of 313 hours for 806 MB of documents on a minicomputer with six Intel 80386 processors. Finally, our own measurements for freeWAIS version 0.202 on a DEC 5000 Model 240 (32 MB memory) with an external disk (Western Scientific) on a SCSI-I bus shows that to index 82.9 MB of our experimental text document database requires 84.1 minutes using ULTRIX V4.2A (Rev. 47) operating system.

8 Conclusion

For dynamic, time critical text document databases, it is important to modify index structures in place, as documents arrive. We have presented a dual structure index strategy to address this problem. Comparing the results presented here with the literature, we have argued that the dual-structure index has better performance than existing implementations with the added bonus of providing incremental updates. The principle source of our improvement is the dynamic division of postings into short and long inverted lists and the application of appropriate data structures to each type of list. Our evaluation is based on using actual data and hardware and simulation of an information retrieval system.

In studying our index, we found a classical trade-off between update time and query time. That is, more time spent incrementally updating the index is repaid with better query performance. We explored algorithms that optimized the time to update and algorithms with optimized query performance and determine various trade-offs between these algorithms. Performance varies by a factor of 4 in the time to build an index and a factor of 20 in query performance.

Another classical trade-off was found between space and time. As the amount of space wasted in storing long inverted lists rises, the query performance to read those inverted lists falls. We described three different methods for allocating additional space on disk to improve query performance and quantitatively describe the trade-off for these methods. In addition, we quantitatively compared overall performance.

For example, the very fastest update performance is possible with poor query performance and poor long list utilization. To get practical query performance and disk utilization for most applications, the time to build an index must be doubled. This increase in time comes from in-place updates that are required to avoid poor disk utilization.

In comparing update performance to query performance, if fast update performance is preferred, we described a policy that offers fast incremental update times with reasonable query performance (the new style with proportional allocation strategy with a constant value of 3.0). Otherwise, if fast query performance is preferred, we presented a policy that offers optimal read performance at a cost of doubling the time to build an index (the whole style with proportional allocation strategy with a constant value of 1.1). We also studied an extent based allocation policy. This policy limits the size of a contiguous region of disk to a fixed maximum amount, so it is easier to implement. However, this feature does have an associated cost (about a 54% increase in query performance for the same disk utilization as the new style).

We also studied the I/O subsystem extensively and determined that the time required to write the bucket data structure to disk is dominated by the subsystem data rate, whereas the time to incrementally update the long lists is dominated by the disk seek time. We quantitatively describe the performance improvements due to speeding up disk or adding more disks. We also determine the performance of updates on an optical disk.

Finally, we believe that further work is required on the dual-structure index since, unfortunately, as the size of the index grows from the addition of more documents, the performance of the index degrades. This implies that we need a strategy to rebalance the division between short and long lists for any number of incremental updates i.e., periodically, as the buckets are read, they can be expanded and written out in a larger region of disk.

Acknowledgements: Thanks to Mendel Rosenblum for discussions on file system mechanisms

related to this paper.

References

[1] Doug Cutting and Jan Pedersen. Optimizations for dynamic inverted index maintenance. In *Proceedings of SIGIR '90*, pages 405–411, 1990.

[2] Samuel DeFazio. Full-text document retrieval benchmark. In Jim Gray, editor, *The Benchmark Handbook for Database and Transaction Processing Systems*, chapter 8. Morgan Kaufmann, second edition, 1993.

[3] Christos Faloutsos and H. V. Jagadish. Hybrid index organizations for text databases. In A. Pirotte, C. Delobel, and G. Gottlob, editors, *Proceedings 3rd International Conference on Extending Database Technology – EDBT '92*, Vienna, 1992. Springer–Verlag.

[4] Christos Faloutsos and H. V. Jagadish. On b-tree indices for skewed distributions. In *Proceedings of 18th International Conference on Very Large Databases*, pages 363–374, Vancouver, British Columbia, Canada, 1992.

[5] William B. Frakes and Ricardo Baeza-Yates. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, 1992.

[6] Donna Harman and Gerald Candela. Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *Journal of the American Society for Information Science*, 41(8):581–589, 1990.

[7] Donald E. Knuth. *The Art of Computer Programming.* Addison-Wesley, Reading, Massachusetts, 1973.

[8] Katia Obraczka, Peter B. Danzig, and Shih-Hao Li. Internet resource discovery services. *IEEE Computer*, 26(9), September 1993.

[9] K. Shoens, A. Luniewski, P. Schwarz, J. Stamos, and J. Thomas. The Rufus system: Information organization for semi-structured data. In *Proceedings of the 19th VLDB Conference*, Dublin, Ireland, 1993.

[10] Stephen Wolfram. *Mathematica.* Addison-Wesley, Redwood City, California, 2nd edition, 1991.

[11] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press, 1949.

[12] Justin Zobel, Alistair Moffat, and Ron Sacks-Davis. An efficient indexing technique for full-text database systems. In *Proceedings of 18th International Conference on Very Large Databases*, Vancouver, 1992.