Efficient Recompression Techniques for Dynamic Full-Text Retrieval Systems

Shmuel T. Klein
Dept. of Math. and Computer Science
Bar-Ilan University, Ramat-Gan 52900, Israel
++(972-3) 531-8865, 531-8408
tomi@bimacs.cs.biu.ac.il

Abstract

An efficient variant of an optimal algorithm is presented, which, in the context of a large dynamic full-text information retrieval system, reorganizes data that has been compressed by an on-the-fly compression method based on LZ77, into a more compact form, without changing the decoding procedure. The algorithm accelerates a known technique based on a reduction to a graph-theoretic problem, by reducing the size of the graph, without affecting the optimality of the solution. The new method can thus effectively improve any dictionary compression scheme using a static encoding method.

1. Introduction and Background

1.1 Compression in Information Retrieval systems

Large Information Retrieval (IR) systems are distributed today on CD-Roms, and even larger systems will fit on a single disk if the main files are being stored in compressed form [14], [2]. Ironically, compression is becoming increasingly important, driven by the advent of new storage capability. For data storage, supply drives demand: the existence of new storage technologies, coupled with improved data capture methods, has greatly increased our appetite to put more data in machine readable form. The largest files in a standard IR system are generally the text itself, but the auxiliary files such as the concordance and/or various files of bitmaps are often as large as the text, if not larger, and should therefore be compressed [5], [6]. We concentrate in this work on the compression of text files, as in [16, 1, 11].

Permission to make digital/hard copies of all or part of this material without fee is granted provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the Association for Computing Machinery, Inc. (ACM). To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or fee.
SIGIR'95 Seattle WA USA© 1995 ACM 0-89791-714-6/95/07 $3.50

Text compression techniques are often divided into statistical methods, such as Huffman coding [10] or arithmetic coding [21], and dictionary methods, based generally on the work of Lempel and Ziv [22], [23]. The statistical methods assign codewords to the elements making up the text, the lengths of these codewords depending on the frequencies of the corresponding elements. Dictionary methods replace variable length substrings of the text by (shorter) pointers to a dictionary in which a collection of such substrings has been stored. Depending on the application and the implementation details, each method can outperform the other, as long as only the *compression savings* are of concern.

While the primary concern is generally to reduce the size of the given file as much as possible, the *time complexity* of the coding routines may also be a relevant factor. For certain applications, such as data transmission over a communication channel, both coding and decoding ought to be fast. For other applications, like the storage of the various files in a large *static* full text IR system, compression and decompression are not symmetrical tasks. Compression is done only once, while building the system, whereas decompression is needed during the processing of every query and directly affects response time. One may thus use extensive and costly preprocessing for compression, provided reasonably fast decompression methods are possible. For large *dynamic* IR systems, however, such as news wires, for which the text is constantly growing, fast and adaptive methods are more appropriate than static ones. Our focus here is on dynamic IR systems.

1.2 Improving dictionary based compression

In LZ77 [22] and its variants, the dictionary is in fact the previously scanned text, and pointers to it are of the form (d, ℓ), where d is an offset (the number of characters from the current location to the previous occurrence of a substring matching the one that starts at the current location), and ℓ is the length of the matching string. There is therefore no need to store an explicit dictionary.

One of the problems of LZ77 is how to locate previous occurrences of substrings in the text. The simple method of scanning the whole text backwards for each processed character might be prohibitively slow. Many alternatives have been suggested, including, among others, the use of binary trees [3], hashing [4, 20] and Patricia trees [9].

The question of how to parse the original text into a sequence of substrings is a problem common to all dictionary based compression techniques. Generally, the parsing is done by a *greedy* method, i.e., at any stage, the longest matching element from the dictionary is sought. A greedy approach is fast, but not necessarily optimal. Because the elements of the dictionary are often overlapping, and particularly for LZ77 variants, where the dictionary is the text itself, a different way of parsing might yield better compression. For example, assume the dictionary consists of the strings $D = \{$`abc`, `ab`, `cdef`, `d`, `de`, `ef`, `f`$\}$ and that the text is $S =$ `abcdef`; assume further that the elements of D are encoded by some fixed-length code, which means that $\lceil \log_2(|D|) \rceil$ bits are used to refer to any of the elements of D; then parsing S by a greedy method, trying to match always the longest available string, would yield `abc-de-f`, requiring 3 codewords, whereas a better partition would be `ab-cdef`, requiring only 2.

The various dictionary compression methods differ also by the way they encode the elements. This is most simply done by a fixed length code, as in the above example. A more involved technique [9] uses a static variable length encoding of the d and ℓ values. Pushing this idea even further, one may use a dynamic variable length code, optimally adapting itself to the frequencies of the occurrences of the different values of d and ℓ: Brent [4] suggests the use of Huffman coding for the (d, ℓ) pairs.

We are concerned here with a way of optimally parsing the text, which may be applied to a process called *recompression*. There are many systems today that offer on-the-fly, very fast, compression of files of any kind. These systems are used to better exploit available disk space, by compressing any file before writing it to the disk. But this is only attractive if the time spent on compression is hardly noticeable, and similarly, decompression must be fast, so that a compressed file may be read without delay. Recompression is useful in a situation where a number of files forming the dynamic IR system have already been compressed by the fast method, and the user wishes now to reorganize the data on his disk into a more compact form. Time is less critical for this reorganization process, but the constraint is that the new encoded form of the recompressed file must be compatible with the original encoding, so that the same decompression method may be used. In other words, a single decoding routine should be able to process a file, regardless of it having been compressed or recompressed.

The method described below has already been mentioned [15, 12], and achieves *optimal* recompression in the sense that once the method for encoding the elements is given, it finds the optimal way of parsing the text into such elements. Obviously, different encoding methods might yield different optimal parsings. Returning to the above example of the dictionary D and text S, if the elements `abc`, `d`, `de`, `ef`, `f`, `ab`, `cdef` of D are encoded respectively by 1, 2, 3, 4, 5, 6 and 6 bits, then the parsing `abc-de-f` would need 9 bits for its encoding, and for the encoding of the parsing `ab-cdef`, 12 bits would be needed. The best parsing, however, for the given codeword lengths, is `abc-d-ef`, which is neither a greedy parsing, nor does it minimize the number of codewords, and requires only 7 bits.

The way to search for the optimal parsing is by reduction to a well-known graph theoretical problem. This approach is, however, not recommended in [15] because of the havy processing involved. In [12], sub-optimal solutions are suggested to improve the execution time. The contribution of this paper is an *efficient* variant of the optimal algorithm: a pruning technique is applied to the graph, which generally reduces the number of both edges and vertices, but still enables the evaluation of an optimal solution for the original graph.

The optimal method and its new variant apply to any dictionary based compression method with static (fixed or variable length) encoding. The elements to be encoded can be of any kind: strings, characters, (d, ℓ) pairs, etc, and any combination thereof. The proposed technique thus improves a very broad range of different methods, many of which have been published in the scientific literature or as patents.

In the next section we mention some simple recompression methods and present the new method and small examples. Examples of encoding functions that have been used and satisfy the required conditions are given in Section 3. Finally, Section 4 states the main

theorems for the analysis.

2. Recompression

Every recompression algorithm corresponds to another tradeoff between the speed of the encoding process and the compression efficiency. Consider a given location in the text, to be encoded by a dictionary compression method. At certain locations, there might be more than one possible choice for the dictionary element to be substituted for the following characters. The algorithm used for scanning the dictionary (linear search, binary search, hashing, etc.) induces an order on the dictionary elements. The range of tradeoff alternatives extends from finding, relative to the ordering at hand, the *first* appropriate element (fastest method, but yielding inferior compression), through considering the k first such elements of the dictionary, for some integer $k > 1$, and choosing the best element among these, up to scanning *all* possible alternatives and selecting the locally optimal element (slower, but giving improved compression).

2.1 Simple recompression methods

For LZ77 and many of its variants, the (d, ℓ) = (*distance, length*) pointers are restricted to $d \leq N$ for some fixed N, that is, a string is considered as recurrent only if its previous occurrence is within a finite window preceding the current location. A simple recompression heuristic is therefore to increase N, which increases the probability of finding a good earlier match. However, the compression performance is not necessarily improved, since $\lceil \log_2 N \rceil$ bits are used to encode d.

In [3, 9, 19], the previous occurrences of the current substring are searched for by means of hashing: the current two (or three) characters are hashed to a location in a hash table, which contains a pointer to the previous occurrence of a couple (or triplet) of characters that hashed to the same location. Since hash functions are not injective, different character pairs or triplets may hash to the same location. It is thus possible that the hash table does not provide a pointer to a previous occurrence, although such an occurrence might exist. There are several ways to use simple recompression in this case. Using a larger hash table will reduce the number of collisions and thereby increase the probability of locating a string if it appeared earlier. Taking this idea a step further, and if enough memory space is available, one could get rid of the hashing altogether, and keep, say, for *every* possible character pair, a pointer to its last occurrence.

In the basic LZ77 algorithm, the *longest* substring is sought which matches the current characters. In the implementations using hashing, this is usually approximated by finding first a matching pair or triplet, and then trying to extend the match as far as possible. This obviously does not guarantee that the longest match will be detected. For instance, if the text that has already been scanned is $T = \ldots$**abcde**$\ldots$**abcx**$\ldots$, and the following characters are **abcde**, then applying hashing to the character pair **ab** will yield, in the better case, a pointer to the last occurrence of **ab**, which can only be extended to form a 3-character match, whereas a 5-character match would have been possible; in the worse case, even that 3-character match will be missed, if another character pair, different from **ab** but yielding the same hash value, has appeared after **abcx** in T.

The compression efficiency can be improved, if not only the last occurrence is remembered, but the k last occurrences, for some constant $k > 1$. For example, one could store pointers to the k last occurrences of each character, using a cyclic list for each. That is, a matrix M of $|\Sigma| \times k$ entries is kept, where $|\Sigma|$ is the size of the alphabet at hand. In addition, an array I of size $|\Sigma|$ is used, with $I[i]$ pointing to the currently used entry in the cyclic list stored in the i-th row of the matrix. Initially, all entries in I are zero. When a character i is encountered in the text, its location is stored in $M[i, I[i]]$, and $I[i]$ is incremented (modulo k). This ensures that up to k previous occurrences can be referenced, and only if more than k occurrences have appeared, the most early ones are overwritten. When a character i is detected, the cyclic list is scanned backwards, and for each of the locations given by the list, the following characters are compared with the characters following the current location; the longest of the up to k matching strings is then used. The drawback of this method is that the same number of memory locations is reserved for each character, whereas in many applications, in particular in natural language texts, certain characters appear much more frequently than others.

Combining this idea of saving multiple references with the hashing approach above, one could store a cyclic list of k elements for each entry of the hash table, or, which is equivalent, have k different hash tables of identical size. If a good hashing function is chosen, the distribution of the hashed addresses will be close to uniform, even if the single character distribution is not. In case there is a strong bias even after hashing, one could use linked lists for each entry of the hash table, thus allowing lists of varying length (up to some predetermined upper limit, induced by the time constraints), without wasting memory locations. However, since hashing functions are non-injective, all the above lists may now contain pointers to different elements.

Consider for example the following scheme: assume the character set consists of the 256 possible 8-bit strings, that a hash table of $2^{14} = 4K$ entries is used, and that hashing is to be applied on character pairs. One could then hash by truncating the least significant bits of each character, i.e.:

$$h(a_7 \cdots a_1 a_0\,,\, b_7 \cdots b_1 b_0) \;=\; a_7 \cdots a_1 b_7 \cdots b_1.$$

If the addresses in the lists are stored in 16-bit words, one could even remove the ambiguity resulting from the hashing, by explicitly storing the truncated bits $a_0 b_0$ in each element of the list. This leaves 14 bits for the address itself, which is equivalent to setting N, the size of the window into which back references should point, to $4K$.

In the next section we consider even better recompression, which recognizes the fact that the longest-matching-string heuristic not necessarily yields an optimal partition.

2.2 Improved optimal recompression

Consider a text string S consisting of a sequence of n characters $S_1 S_2 \cdots S_n$, each character S_i belonging to a fixed alphabet Σ. Substrings of S are referenced by their limiting indices, i.e., $S_i \cdots S_j$ is the substring starting at the i-th character in S, up to and including the j-th character. We wish to compress S by means of a dictionary D, which is a set of character strings $\{\sigma_1, \sigma_2, \ldots\}$, with $\sigma_i \in \Sigma^+$. The dictionary may be explicitly given and finite, as in the example in the introduction, or it may be potentially infinite, e.g., for the Lempel-Ziv variants, where any previously occurring string can be referenced.

The compression process consists of two independent phases: parsing and encoding. In the *parsing* phase, the string S is broken into a sequence of consecutive sub-strings, each belonging to the dictionary D, i.e., an increasing sequence of indices $i_0 = 0, i_1, i_2, \ldots$ is found, such that

$$S \;=\; S_1 S_2 \cdots S_n \;=\; S_1 \cdots S_{i_1}\, S_{i_1+1} \cdots S_{i_2} \;\cdots,$$

with $S_{i_j+1} \cdots S_{i_{j+1}} \in D$ for $j = 0, 1, \ldots$. One way to assure that at least one such parsing exists is to force the dictionary D to include each of the individual characters of Σ. The second phase is based on an *encoding* function $\lambda : D \longrightarrow \{0,1\}^*$, that assigns to each element of the dictionary a binary string, called its encoding. The assumption on λ is that it produces a code which is uniquely decipherable (UD). This is most easily obtained by a fixed length code, but such a code is only possible for a finite dictionary, and even then it is only efficient from the compression point of view if the distribution of the occurrences of the elements of D in S is nearly uniform. Compression can often be improved by the use of variable-length codes, assigning shorter codewords to elements with higher probability of occurrence. A sufficient condition for a code being UD is to choose it as a prefix code (see Even [8]).

The problem is the following: given the dictionary D and the encoding function λ, we are looking for the optimal partition of the text string S, i.e., the sequence of indices $i_1, i_2, \ldots$ is sought, that minimizes $\sum_{j \geq 0} |\lambda(S_{i_j+1} \cdots S_{i_{j+1}})|$.

To solve the problem, a directed, labeled graph $G = (V, E)$ is defined for the given text S. The set of vertices is $V = \{1, 2, \ldots, n, n+1\}$, with vertex i corresponding to the character S_i for $i \leq n$, and $n+1$ corresponding to the end of the text; E is the set of directed edges: an ordered pair (i, j), with $i < j$, belongs to E if and only if the corresponding substring of the text, that is, the sequence of characters $S_i \cdots S_{j-1}$, can be encoded as a single unit. In other words, the sequence $S_i \cdots S_{j-1}$ must be a member of the dictionary, or more specifically for LZ77, if $j > i+1$, the string $S_i \cdots S_{j-1}$ must have appeared earlier in the text. The label L_{ij} is defined for every edge $(i, j) \in E$ as $|\lambda(S_i \cdots S_{j-1})|$, the number of bits necessary to encode the corresponding member of the dictionary, for the given encoding scheme at hand. The problem of finding the optimal parsing of the text, relative to the given dictionary and the given encoding scheme, therefore reduces to the well-known problem of finding the shortest path in G from vertex 1 to vertex $n+1$.

Dijkstra's [7] algorithm may be used to find the shortest path. Its worst case complexity is, depending on the data structures used, $O(|V|^2)$, or $O(|E| \log |V|)$ (see [18]), which would be particularly disturbing for our intended application. However, in our case the directed graph contains no cycles, since all edges are of the form (i, j) with $i < j$. Thus by a simple dynamic coding method, the shortest path can be found in $O(|E|)$. Nevertheless, when the text includes long runs of repeated characters (like strings of zeros or blanks), the number of possibilities to parse these runs, and hence the number of edges, is quadratic in the number of vertices. This motivated the search for sub-optimal alternatives in [12].

We suggest here to adhere to the optimal parsing, and to circumvent the worst case behavior by combining the shortest path algorithm with a *pruning* method intended to eliminate a priori such parts of the graph, that cannot possibly be part of an optimal path. The pruning process may be applied in all cases for which

the labeling function L satisfies the triangle inequality,

$$L_{ij} \leq L_{ik} + L_{kj} \text{ for all } i,k,j \text{ such that } i < k < j,$$

which holds for many practical encoding schemes (see next section for examples).

The set of edges E is constructed dynamically by the algorithm itself. We start with $E = \emptyset$, and adjoin, in order, the edges emanating from vertices $1, 2, \ldots$, unless they fail to pass the following test. When a vertex i is reached, consider the set of its predecessors $\text{Pred}(i) = \{j \mid (j,i) \in E\}$. We now scan the substrings of the text starting at S_i. Suppose that the substring $S_i \cdots S_{j-1}$ is a member of the dictionary, so that the pair (i,j) is a candidate to be adjoined to E. Before adding this edge, check if it is possible to reach vertex j directly from every vertex in $\text{Pred}(i)$, without passing through vertex i. If so, then there is no need to add the edge (i,j) to E, since, because of the triangle inequality, there is no loss in taking the direct edge from the element of $\text{Pred}(i)$ to vertex j. However, if there is even one element in $\text{Pred}(i)$ that has no direct edge to j, then (i,j) must be added to E.

If, after having checked all the edges emanating from vertex i, none of these have been adjoined to E, then there is no need to keep the vertex i in the graph, since it obviously cannot be part of an optimal path from 1 to $n+1$. Thus all the incoming edges on vertex i may be pruned from the graph, and i itself may also be eliminated.

The formal definition of the algorithm is given on the opposite column.

The algorithm seems non-symmetric with regard to the predecessors and successors of a vertex. This is because the vertices are scanned sequentially. Therefore, when processing vertex i, $\text{Pred}(i)$ is already defined, but not yet $\text{Succ}(i) = \{j \mid (i,j) \in E\}$, the set of i's successors. The set of the *potential* successors of i, $\text{Succ\_Candidates}(i)$, is defined as the set of those vertices to which there would have been a direct edge if no pruning were used. Some of these edges might ultimately not be adjoined to the graph. Others might in a first stage be added, but might later be deleted.

Note that the triangle inequality is a sufficient condition for reaching an optimal solution with the improved algorithm, but when the condition does not hold for every triple (i,j,k), one can easily adapt the algorithm to deal also with these cases: replace the test **if** $(k,j) \notin E$ in the inner loop by

$$\textbf{if } (k,j) \notin E \textbf{ or } L_{kj} > L_{ki} + L_{ij}.$$

In other words, even if we can reach j from all the predecessors k of i, we still might have to add the edge (i,j) to the graph.

```
Algorithm Prune
{
    E ⟵ ∅
    V ⟵ {1, ..., n, n+1}
    for i ⟵ 1 to n
    {
        Pred(i) ⟵ {k | (k,i) ∈ E}
        Succ_Candidates(i) ⟵ {j | S_i ··· S_{j-1} ∈ D}
        added_edge ⟵ FALSE
        for all j ∈ Succ_Candidates(i)
        {
            all_connected ⟵ TRUE
            for all k ∈ Pred(i)
            {
                if (k,j) ∉ E
                then all_connected ⟵ FALSE
            }
            if not all_connected then
            {
                E ⟵ E ∪ {(i,j)}
                added_edge ⟵ TRUE
            }
        }
        if not added_edge then
        {
            V ⟵ V \ {i}
            E ⟵ E \ Pred(i)
        }
    }
}
```

Figure 1 displays a small example of a graph, corresponding to the text **abbaabbabab**, including all the vertices and edges. We now assume that LZ77 is used. The edges connecting vertices i to $i+1$, for $i = 1, \ldots, n$, are labeled by the character S_i. Note that the example clearly displays the main problem with long recurring strings: if $S_i \cdots S_{j-1}$ did occur earlier, so did also all its substrings $S_k \cdots S_\ell$, for $i \leq k \leq \ell < j$, therefore the the corresponding sub-graph with vertices $\{i, \ldots, j\}$ is a full graph. For example, the sub-graph on vertices $\{5,6,7,8,9\}$ corresponds to the string **abba**, and $\{9,10,11,12\}$ corresponds to the string **bab**. If such recurring strings form a major part of the text, the number of edges might be $\Omega(|V|^2)$.

After having applied the pruning algorithm, many edges may have been deleted, as well as some of the vertices. Figure 2 depicts the graph obtained for the same text as for Figure 1, but with the use of the pruning algorithm. The character corresponding to the transition from vertex i to $i+1$ is indicated, even if the corresponding edge has been deleted.

The example also indicates how the algorithm could

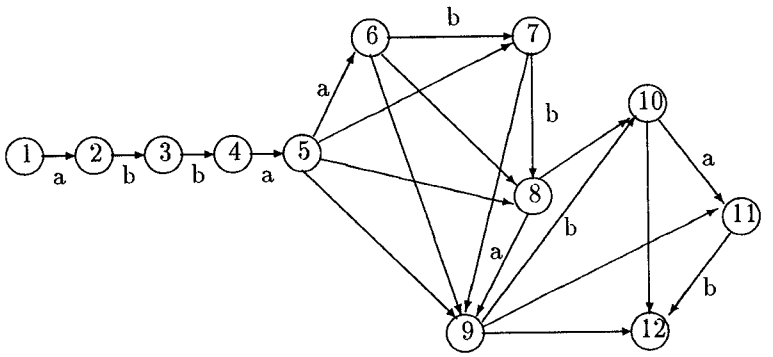

Figure 1: Original graph corresponding to text **abbaabbabab**

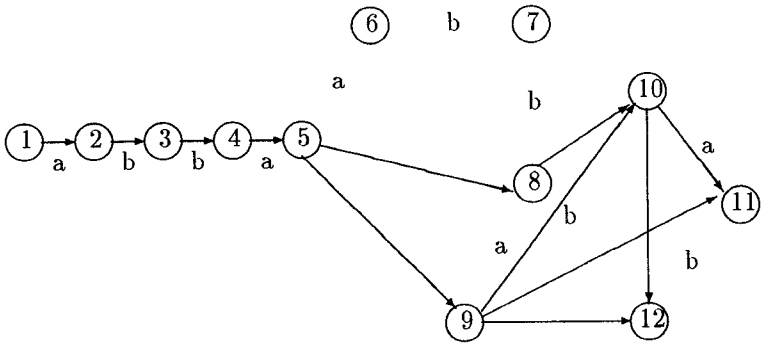

Figure 2: Graph for text **abbaabbabab** after pruning

be improved. Note that since all the successors of 10 are also successors of 9, the triangle inequality in fact implies that the edge (9,10) could also be eliminated. The reason this is not done in the algorithm is because the loop with running index k, passing over the predecessors, is internal to the loop with running index j, passing over the potential successors. Therefore, when (9,10) is adjoined, the set of the successors of 10 is not known yet. An additional loop checking also such cases might further improve the algorithm, but is not necessarily justified.

The routine evaluating then the shortest path from 1 to $n+1$ uses an array SPL(i), for storing the Shortest Path Length from 1 to i. In iteration i, the values of SPL(j) for $j < i$ are already known. The algorithm now scans only those vertices and edges that remain after the pruning process.

Since for each i and j, the label L_{ji} is referenced exactly once by the algorithm, its time complexity is clearly $O(|E|)$.

```
Shortest path
{
    SPL(1) ⟵ 0
    for i ⟵ 2 to n+1
        if i ∈ V then
        {
            SPL(i) ⟵ ∞
            for all j ∈ Pred(i)
            {
                t ⟵ SPL(j) + L_ji
                if SPL(i) < t then SPL(i) ⟵ t
            }
        }
}
```

3. Encoding function examples

This section brings examples of encoding functions λ that have been proposed and are used in commercial compression systems. We show that they obey the triangle inequality, which is a sufficient condition for applying the above pruning algorithm.

The first example, based on [19], is a variant of LZ77, known as LZSS [17], using hashing on character pairs to locate (the beginning of) recurrent strings, like in [20]. The output of the compression process is thus a sequence of elements, each being either a single (uncompressed) character, or an offset-length pair (d, ℓ). The elements are identified by a flag bit, so that a single character is encoded by a zero, followed by the 8-bit ASCII representation of the character, and the encoding of each (d, ℓ) pair starts with a 1. The sets of possible offsets and lengths are split into classes as follows: let $B_m(n)$ denote the standard m-bit binary representation of n (with leading zeros if necessary), then, denoting the encoding scheme by λ_S:

$$\lambda_S(\text{offset } d) = \begin{cases} 1B_7(d) & \text{if } d \le 127 \\ 0B_{11}(d) & \text{if } 127 < d \le 2047 \end{cases}$$

$$\lambda_S(\text{length } \ell) = \begin{cases} B_2(\ell-2) & \text{if } 2 \le \ell \le 4 \\ 11B_2(\ell-5) & \text{if } 5 \le \ell \le 7 \\ (1111)^{\lceil(\ell-7)/15\rceil} B_4((\ell-8) \bmod 15) & \\ \qquad \text{if} \quad \ell \ge 8 & \end{cases}$$

Including the flag-bit, each offset is thus encoded by 9 or 13 bits, and the number of bits used to encode the length ℓ is $2\lceil\frac{\ell}{4}\rceil$ for $\ell < 8$ and it is $4\lceil\frac{\ell+8}{15}\rceil$ for $\ell \ge 8$. It is of course wasteful to use an encoding of linearly growing length for the values of ℓ, but the decoding speed is enhanced, since only half-byte blocks are processed (except for $\ell \le 4$).

Theorem 1 *The function λ_S satisfies the triangle inequality.*

Proof: Let E_1 and E_2 be two consecutive elements encoded by the algorithm, where E_i may be either a single character, or a string of characters encoded by an (offset, length) pair (d, ℓ). Denote by E the concatenation of E_1 with E_2, and assume that E may be encoded as a single element. Let $L(x)$ be the function giving the length, in bits, of the encoding $\lambda_S(x)$, where, as above, we shall apply L to both the offset or the length part, or even to an element E_i. We have to show that $L(E) \le L(E_1) + L(E_2)$.

Case 1: Both E_i are single characters, then $L(E_1) = L(E_2) = 9$. But E is a string of two characters and will be encoded by an (d, ℓ) pair. The *offset* part is encoded by at most 13 bits, the *length* part by exactly two bits (since $\ell = 2$). Thus $L(E) \le 13 + 2 < 9 + 9$.

Case 2: One of the E_i is a single character, the other a string encoded by (d, ℓ). Then there exists a d' such that E is encoded by $(d', \ell + 1)$. Thus

$$L(E_1) + L(E_2) - L(E) \ge 9 + (9 + L(\ell)) - (13 + L(\ell + 1)) \ge 1,$$

since the difference in the lengths of the encodings of consecutive lengths ℓ and $\ell + 1$ never exceeds 4 bits.

Case 3: Both E_i are strings, encoded by (d_i, ℓ_i), respectively. Then there exists a d' such that E is encoded by $(d', \ell_1 + \ell_2)$.

If both ℓ_1 and ℓ_2 are smaller than 8, then they are encoded together by at least 4 bits, but $\ell_1 + \ell_2$ is at most 14, so it is encoded by at most 8 bits. If, say, ℓ_1 is smaller than 8, but ℓ_2 is not, then they are encoded together by at least $L(\ell_2) + 2$ bits, but $\ell_1 + \ell_2$ is at most $\ell_2 + 7$, so it is encoded by at most $L(\ell_2) + 4$ bits. If both ℓ_i are larger than 7, then

$$\begin{aligned} L(\ell_1) + L(\ell_2) &\ge 4\left(\tfrac{\ell_1+8}{15} + \tfrac{\ell_2+8}{15}\right) \qquad \text{and} \\ L(\ell_1 + \ell_2) &\le 4\left(\tfrac{\ell_1+\ell_2+8}{15} + 1\right), \end{aligned}$$

so that $L(\ell_1 + \ell_2) - (L(\ell_1) + L(\ell_2)) \le 4 - \frac{32}{15} < 2$. Thus for all values of ℓ_1 and ℓ_2, $L(\ell_1 + \ell_2)$ exceeds $L(\ell_1) + L(\ell_2)$ by at most 4 bits. Therefore

$$\begin{aligned} & L(E_1) + L(E_2) - L(E) \\ \ge\ & (9 + L(\ell_1)) + (9 + L(\ell_2)) - (13 + L(\ell_1 + \ell_2)) \\ \ge\ & 1. \end{aligned}$$

∎

The second example comes from the on-the-fly compression routine recently included in a popular operating system. It is again based on [20], but uses simpler hashing and a different encoding scheme λ_M. Single characters are again encoded by 9 bits, and the sets of offsets and lengths are encoded as follows:

$$\lambda_M(\text{offset } d) = \begin{cases} 1B_6(d-1) & \text{if } 1 \le d \le 64 \\ 01B_8(d-65) & \text{if } 64 < d \le 320 \\ 11B_{11}(d-321) & \text{if } 320 < d \le 2368 \end{cases}$$

$$\lambda_M(\text{length } \ell) = \begin{cases} 0 & \text{if } \ell = 2 \\ 1^{j+1}\ 0\ B_j(\ell - 2 - 2^j) & \\ \qquad \text{if } 2^j \le \ell - 2 < 2^{j+1}, & \\ \qquad \text{for } j = 0, 1, 2, \ldots & \end{cases}$$

Offsets are thus encoded by 8, 11 or 14 bits, and the number of bits used to encode the lengths ℓ is 1 for $\ell = 2$ and $2\lceil\log_2(\ell - 1)\rceil$ for $\ell > 2$.

Theorem 2 *The function λ_M satisfies the triangle inequality.*

Proof: We use the same notations as for Theorem 1, $L(x)$ standing now for $|\lambda_M(x)|$, and consider the same three cases.

Case 1: $L(E_1) + L(E_2) = 9 + 9 > 14 + 1 \geq L(d, 2) = L(E)$.

Case 2: $L(E_1)+L(E_2)-L(E) \geq 9+(8+L(\ell))-(14+L(\ell+1)) \geq 1$, since the difference in the lengths of the encodings of consecutive lengths ℓ and $\ell+1$ never exceeds 2 bits.

Case 3: For the case $\ell_1 = \ell_2 = 2$, we have $L(\text{length } 2) + L(\text{length } 2) - L(\text{length } 4) = 1 + 1 - 4 = -2$. If, say, $\ell_1 = 2$ and $\ell_2 > 2$, then we note that $L(\ell_2+2)$ exceeds $L(\ell_2)$ by at most 2 bits, thus $L(\text{length } 2) + L(\ell_2) - L(\ell_2 + 2) \geq 1 - 2 = -1$. If both ℓ_i are larger than 2, then using the fact that the logarithmic functions are sub-additive, we get

$$\begin{aligned} & L(\ell_1) + L(\ell_2) - L(\ell_1 + \ell_2) \\ \geq\ & 2\left(\log_2(\ell_1 - 1) + \log_2(\ell_2 - 1) \right. \\ & \qquad \left. - (\log_2(\ell_1 + \ell_2 - 1) + 1)\right) \\ \geq\ & 2\left(\log_2(\ell_1 + \ell_2 - 2) - \log_2(\ell_1 + \ell_2 - 1) + 1\right) \\ =\ & 2\left(\log_2\left(1 - \tfrac{1}{\ell_1+\ell_2-1}\right) + 1\right) \geq 1.3562, \end{aligned}$$

the last inequality following from the fact that $\log_2(1 - \frac{1}{n})$ is increasing with n, and we consider here values $n \geq 5$. Thus for all values of ℓ_1 and ℓ_2, $L(\ell_1 + \ell_2)$ exceeds $L(\ell_1) + L(\ell_2)$ by at most 2 bits. Therefore

$$\begin{aligned} & L(E_1) + L(E_2) - L(E) \\ \geq\ & (8 + L(\ell_1)) + (8 + L(\ell_2)) - (14 + L(\ell_1 + \ell_2)) \\ \geq\ & 0. \end{aligned}$$

∎

Remark: A newer variant of the second example extends the size of the window into which a back-reference may point. The last range of λ_M(offset d) is then encoded by $11B_{12}(d - 321)$ if $320 < d \leq 4416$, so that offsets may be encoded by up to 15 bits. This, however, affects the triangle inequality. In the special case when two consecutive strings **ab** and **cd** appeared earlier at distances d_1 and d_2, both ≤ 64, but the concatenated string **abcd** appeared earlier at a distance $d_3 > 320$, encoding the string **abcd** requires $L(d_3, 4) = 15 + 4$ bits, which is larger than $L(d_1, 2)+L(d_2, 2) = (8+1)+(8+1)$ bits, needed to encode the pairs **ab** and **cd** individually. As mentioned above, the Prune algorithm may be adapted to deal with such cases too.

4. Analysis

One problem in the implementation is to have an easy way to keep track of the predecessors of each vertex. If no pruning is used, every preceding vertex may be a predecessor of the current vertex i, but in fact, when scanning backwards, we may stop as soon as a vertex is found which is not connected to i, since the predecessors of a vertex must immediately precede it. When the pruning algorithm is applied, the fact that the predecessors of a vertex immediately precede it is not necessarily true. Nevertheless, the predecessors form a contiguous block, so that while scanning backwards, once at least one predecessor j of i has been detected, we may continue sequentially to $j-1$, $j-2$, etc., and stop as soon as the first vertex $j-k$ is found, with $k > 0$, which is not a predecessor of i.

For the theorems below, we need to differentiate between two kinds of predecessors and successors of the vertices. There are two kinds of edges missing from the graph: those that have not been added at all, and those that have been added, but were deleted later. Define Pred'(i) as the set of vertices which were predecessors of i at some stage of the algorithm, and Succ'(i) as the set of vertices which were successors of i at some stage of the algorithm. The sets $\text{Pred}(i) = \{j \mid (j,i) \in E\}$ and $\text{Succ}(i) = \{j \mid (i,j) \in E\}$ defined earlier refer to the vertices which are predecessors or successors of i even after the algorithm has completed its task. Clearly, $\text{Pred}(i) \subseteq \text{Pred'}(i)$ and $\text{Succ}(i) \subseteq \text{Succ'}(i)$.

Theorem 3 *If pruning is used, then for each node x,* Succ'(x) *consists of consecutive elements, i.e.:*

$$|\text{Succ'}(x)| = k \quad \rightarrow \quad \exists j \geq 0 \quad \text{such that} \quad \text{Succ'}(x) = \{x + j + 1, x + j + 2, \ldots, x + j + k\}$$

Theorem 4 *If pruning is used, then for each node $x > 0$,* Pred'(x) *consists of consecutive elements, i.e.:*

$$|\text{Pred'}(x)| = k \quad \rightarrow \quad \exists j \geq 0 \quad \text{such that} \quad \text{Pred'}(x) = \{x - j - k, \ldots, x - j - 2, x - j - 1\}.$$

The proofs are non-trivial inductions on x. They are omitted for lack of space and will appear in the full paper.

The fact that the last theorem is about Pred'(x) and not about Pred(x) is no real restriction, since anyway, the stage where the algorithm has to scan the predecessors of a vertex x is prior to the moment where any edges incident on x may be deleted.

5. Conclusion

The new technique improves the running time of the optimal algorithm, and may thus efficiently enhance almost all of the dictionary based encoding methods that have been suggested. Our experimental results show an additional 20%–40% savings obtained by the optimal recompression algorithm, relative to the on-the-fly methods. Relative to implementations of the greedy longest-match heuristic, the improvement was, on most data, only a modest one of a few percent, as predicted by [13]. However, even a minor improvement might be worthwhile in certain applications, and has certainly theoretical value, since one can show that it cannot be further improved under our constraints of a predetermined decoding procedure. For our application to dynamic full-text IR systems, the Prune algorithm may permit to replace sub-optimal compression heuristics by an optimal method, without major loss in processing speed.

References

[1] **Bell T.C., Cleary J.G., Witten I.A.,** *Text Compression*, Prentice Hall, Englewood Cliffs, NJ (1990).

[2] **Bell T.C., Moffat A., Nevill-Manning C.G., Witten I.H., Zobel J.,** Data compression in full-text retrieval systems, *J. Amer. Soc. of Inf. Sci.* **44** (1993) 508–531.

[3] **Bell T.C.,** Better OPM/L text compression, *IEEE Trans. on Communications* **COM–34** (1986) 1176–1182.

[4] **Brent R.P.,** A linear algorithm for data compression, *The Australian Computer Journal* **19** (1987) 64–68.

[5] **Choueka Y., Fraenkel A.S., Klein S.T.,** Compression of Concordances in Full-Text Retrieval Systems, *Proc. 11-th ACM-SIGIR Conf.*, Grenoble (1988) 597–612.

[6] **Choueka Y., Fraenkel A.S., Klein S.T., Segal E.,** Improved Hierarchical Bit-Vector Compression in Document Retrieval Systems, *Proc. 9-th ACM-SIGIR Conf.*, Pisa (1986) 88–97.

[7] **Dijkstra E.W.,** A note on two problems in connexion with graphs, *Numerische Mathematik* **1** (1959) 269–271.

[8] **Even S.,** *Graph Algorithms*, Computer Science Press (1979).

[9] **Fiala E,R., Greene D.H.,** Data compression with finite windows, *Comm. of the ACM* **32** (1989) 490–505.

[10] **Huffman D.,** A method for the construction of minimum redundancy codes, *Proc. of the IRE* **40** (1952) 1098–1101.

[11] **Lelewer D.A., Hirschberg D.S.,** Data compression, *ACM Computing Surveys* **19** (1987) 261–296.

[12] **Katajainen J., Raita T.,** An approximation algorithm for space-optimal encoding of a text, *The Computer Journal* **32** (1989) 228–237.

[13] **Katajainen J., Raita T.,** An analysis of the longest match and the greedy heuristics in text encoding, *J. ACM* **39** (1992) 281–294.

[14] **Klein S.T., Bookstein A., Deerwester S.,** Storing Text Retrieval Systems on CD-ROM: Compression and Encryption Considerations, *ACM Trans. on Information Systems* **7** (1989) 230–245.

[15] **Schuegraf E.J., Heaps H.S.,** A comparison of algorithms for data base compression by use of fragments as language elements, *Information Stor. and Retr.* **10** (1974) 309–319.

[16] **Storer J.A.,** *Data Compression: Methods and Theory,* Computer Science Press, Rockville, Maryland (1988).

[17] **Storer J.A., Szymanski, T.G.,** Data compression via textual substitution, *J. ACM* **29** (1982) 928–951.

[18] **Tarjan R.E.,** *Data Structures and Network Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia (1983).

[19] **Whiting D.L., George G.A., Ivey G.E.,** Data Compression Apparatus and Method, U.S. Patent 5,126,739 (1992).

[20] **Williams R.N.,** An extremely fast Ziv-Lempel data compression algorithm, *Proc. Data Compression Conference DCC–91*, Snowbird, Utah (1991) 362–371.

[21] **Witten I.H, Neal R.M., Cleary J.G.,** Arithmetic coding for data compression, *Comm. ACM* **30** (1987) 520–540.

[22] **Ziv J., Lempel A.,** A universal algorithm for sequential data compression, *IEEE Trans. on Inf. Th.* **IT–23** (1977) 337–343.

[23] **Ziv J., Lempel A.,** Compression of individual sequences via variable-rate coding, *IEEE Trans. on Inf. Th.* **IT–24** (1978) 530–536.