Computing Longest Previous Factor in linear time and applications

Maxime Crochemore a,b,1, Lucian Ilie c,*2

a Department of Computer Science, King’s College London, London WC2R 2LS, UK
b Institut Gaspard-Monge, Université Paris-Est, F-77454 Marne-la-Vallée Cedex 2, France
c Department of Computer Science, University of Western Ontario, N6A 5B7, London, Ontario, Canada

Received 27 July 2007; received in revised form 16 October 2007; accepted 17 October 2007
Available online 23 October 2007

Communicated by L.A. Hemaspaandra

Abstract

We give two optimal linear-time algorithms for computing the Longest Previous Factor (LPF) array corresponding to a string \( w \). For any position \( i \) in \( w \), \( \text{LPF}[i] \) gives the length of the longest factor of \( w \) starting at position \( i \) that occurs previously in \( w \). Several properties and applications of LPF are investigated. They include computing the Lempel–Ziv factorization of a string and detecting all repetitions (runs) in a string in linear time independently of the integer alphabet size.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Design of algorithms; Analysis of algorithms; Strings; Suffix array; Longest common prefix; Longest previous factor; Lempel–Ziv factorization; Repetitions; Runs

1. Introduction

Given a string \( w \), we introduce the Longest Previous Factor (LPF) array defined as follows. For any position \( i \) in \( w \), \( \text{LPF}[i] \) gives the length of the longest factor of \( w \) starting at position \( i \) that occurs previously in \( w \). Formally, if \( w[i] \) denotes the \( i \)th letter of \( w \) and \( w[i..j] \) is the factor \( w[i]w[i + 1]...w[j] \), then

\[
\text{LPF}[i] = \max\{\ell \mid w[i..i + \ell - 1] \text{ is a factor of } w[0..i + \ell - 2] \cup \{0\}\}.
\]

We give two linear-time (optimal) algorithms for computing LPF using suffix arrays. The first uses no additional information whereas the second uses the longest common prefix array which is often part of the suffix array data structure. Previously such algorithms involved computing the suffix trees, which are more complex and take a lot of space. Also, a logarithmic factor of the size of the alphabet often appears in the complexity. Our algorithms use suffix arrays, are much simpler, and their complexity is alphabet independent.

One important application is computing the Lempel–Ziv factorization [14]. Recently Abouelhoda et al. [1] gave a suffix-array-based algorithm for computing Lempel–Ziv factorization. However, their algorithm is...
essentially a simulation of the suffix tree using the suffix array. The description in [1] is very brief but it seems that their approach can be used to achieve similar goals with ours, nevertheless in a significantly more complicated way.

Simultaneously and independently of our work, Chen et al. [2] gave an algorithm that is similar with ours, nevertheless in a significantly more complicated way. The idea of the algorithm is as follows. For any position \( i \), the longest factor starting at \( i \) that occurs also to the left of \( i \) in \( w \) is the longest common prefix between the suffix \( suf_i \) and the suffixes starting to the left of \( i \) in \( w \), that is, \( suf_j \), \( 0 \leq j < i \). However, given \( SA \), we need only consider those which are closest to \( suf_i \) in \( SA \). We shall therefore compute, for each \( i \), the closest positions in \( SA \) that are smaller than \( i \); in most cases there will be two such positions, one before and one after \( i \) in \( SA \). Denote them by \( \text{prev}_<[SA[i]] \) and \( \text{prev}_>[SA[i]] \), respectively. If one of them does not exist, then we assign the value \(-1\); see columns 5 and 6 in Fig. 1. Rephrasing the above, \( LPF[i] \) is obtained as the length of the longest common prefix between \( suf_i \) and either \( suf_{\text{prev}_<[i]} \) or \( suf_{\text{prev}_>[i]} \), whichever is longer.

After \( \text{prev}_< \) and \( \text{prev}_> \) are found, \( LPF \) is computed for all values of \( i \) in increasing order, using the property that \( LPF[i] \geq LPF[i-1] - 1 \). Thus, we already know that \( w[i..i + LPF[i-1] - 2] \) occurred to the left of \( i \) and need only try to extend it. A problem appears because we do not know whether we should compare \( suf_i \) to \( suf_{\text{prev}_<[i]} \) or \( suf_{\text{prev}_>[i]} \). We shall therefore compute two arrays, \( LPF_{<}[0..n-1] \) and \( LPF_{>}[0..n-1] \); they have the same meaning as \( LPF \) except that they consider only positions corresponding to suffixes lexicographically smaller, resp. larger, than \( suf_i \). Formally, we have

3 Note that a suffix-tree-based algorithm would compute the leftmost such position in the string whereas our algorithm might produce a different one. For instance, in our example, \( \text{PrevOcc}[12] = 10 \) but the leftmost occurrence of ab starts at 0.
Theorem 1. Given a string of length \( n \) over an integer alphabet, the LPF and \( \text{PrevOcc} \) arrays can be computed in time and space \( O(n) \).

4. An algorithm using LCP

Our second algorithm for computing LPF uses the LCP array. Its advantage over the previous algorithm is that it processes the suffix array in one pass and requires less memory space. The idea is similar to the one above. Assuming we know the longest common prefixes between \( \text{suf}_{i} \) and either \( \text{suf}_{\text{prev},i} \) or \( \text{suf}_{\text{prev},i} \), it is enough to take the maximum of the two.

Using this idea, we give a space-efficient algorithm for computing LPF. For a better understanding, it is useful to arrange the SA and LCP arrays in a graph. The vertices are labelled by the SA values and the edges by the LCP values. The vertices are arranged in the left-to-right order corresponding to their order in SA and are placed at a height corresponding to their starting position in the string. In other words, if \( \text{SA}[i] = j \), then the vertex labelled \( j \) is plotted with abscissa \( i \) and ordinate \( j \). An example is shown by the graph in Fig. 3(i) consisting of the solid edges only.

Now consider the vertices in decreasing order of their labels, that is, vertices that are highest in the graph (“peaks”) are considered first. For vertex 13, the two adjacent edges are labelled 0 and 1, corresponding to the longest common prefixes of \( \text{suf}_{13} \) with \( \text{suf}_{\text{prev},13} = \text{suf}_{4} \) and \( \text{suf}_{\text{prev},13} = \text{suf}_{7} \), respectively. Therefore, the maximum of the two gives \( \text{LPF}[13] = 1 \). On the other hand, the minimum of the two gives the longest common prefix of \( \text{suf}_{4} \) and \( \text{suf}_{7} \); we remove the vertex 13 and add a (dotted) edge between 4 and 7, labelled 0. Note that \( \text{prev}_{4} = 7 \) and \( \text{prev}_{7} = 2 \), so this property is an invariant of the graph. Next we consider the vertex 12 and so on. Fig. 3(ii) shows the graph after having considered the vertices 13, 12, and 11.

It is clear that we need not consider the vertices in this order. For instance, we can compute \( \text{LPF}[6] \) right away. Precisely, any vertex which is a “peak” in our graph can have its LPF value computed. In the algorithm in Fig. 4 we consider the vertices in the order they appear in the SA (that is, left to right in the graph) and use a stack to store unprocessed vertices. The vertices in the stack will be in increasing order from bottom to top. Whenever a vertex with a smaller label than the one on top of the stack is encountered, the top of the stack is processed and popped. For instance, vertex 4 has to wait for 13 and 7 to be processed. In order to process all vertices uniformly, we assume a last (virtual) vertex with label \( n \) and height \(-1\). The stack contains pairs
Fig. 3. (i) Solid edges form the graph representing $SA$ and $LCP$ for the text $abbaabbaaabab$; dotted edges show the (conceptual) transformation of the graph during the algorithm COMPUTE LPF USING LCP. (ii) The graph after the vertices 13, 12, and 11 were considered.

**Algorithm COMPUTE LPF USING LCP**

1. $SA[n] \leftarrow -1; LCP[n] \leftarrow 0$
2. PUSH((0, $SA[0]$), $\delta$)
3. for $i$ from 1 to $n$ do
4. $lcp \leftarrow LCP[i]$
5. while ($\delta \neq \emptyset$ and ($SA[i] < \text{TOP}(\delta).\text{pos}$)) do
6. $LPF[\text{TOP}(\delta).\text{pos}] \leftarrow \max(\text{TOP}(\delta).\text{len}, \text{lcp})$
7. $\text{lcp} \leftarrow \min(\text{TOP}(\delta).\text{len}, \text{lcp})$
8. $v \leftarrow \text{TOP}(\delta)$
9. POP($\delta$)
10. if ($LPF[v.\text{pos}] = 0$) then
11. $\text{PrevOcc}[v.\text{pos}] \leftarrow -1$
12. else if ($v.\text{len} > \text{lcp}$) then
13. $\text{PrevOcc}[v.\text{pos}] \leftarrow \text{TOP}(\delta).\text{pos}$
14. else $\text{PrevOcc}[v.\text{pos}] \leftarrow SA[i]$
15. if ($i < n$) then PUSH((lcp, $SA[i]$), $\delta$)
16. return $LPF$

Fig. 4. Algorithm for computing LPF using LCP.

of the form $(x.\text{len}, x.\text{pos})$, where $x.\text{pos}$ is a position (in $SA$) and $x.\text{len}$ stores the longest common prefix between $suf_\text{x.\text{pos}}$ and the suffix corresponding to the node right below $x$ in the stack (or 0 if none). For instance, when the vertex 7 is considered, 13 is processed and removed from the stack. The top of the stack becomes 4 and 7 is then pushed on top of it: TOP($\delta$).len $\leftarrow 0$ and TOP($\delta$).pos $\leftarrow 7$.

The correctness of the algorithm follows from the above discussion. It runs in $O(n)$ time because each element of $SA$ is pushed only once on to the stack. Also, [10] gives a very simple linear time algorithm to compute the LCP array. The spaced used by the stack is at most $n$ pairs of integers, which is reached for the string $a^{n-1}b$. However, the expected size of the stack is much less.

There is an interesting consequence of the above discussion, namely that the array $LPF$ is a permutation of $LCP$. This is shown by analyzing the graph in Fig. 3(i). The labels of the solid edges form the LCP array. When removing a “peak” from the graph, the maximum of the two labels of the adjacent edges becomes an $LPF$ value whereas the minimum becomes the label of the newly formed (dotted) edge. Each value will eventually become an $LPF$ value which proves the statement.

**Proposition 1.** $LPF$ is a permutation of $LCP$.

**Remark 1.** Note that [1] suggests a bottom-up computation of the $LPF$ array on the lcp-interval tree (isomorphic with the suffix tree) and therefore leads to a similar result with ours. However, our approach is much simpler.

5. Application 1: Computing the Lempel–Ziv factorization

The Lempel–Ziv factorization of $w$ [14] is the decomposition $w = u_0u_1 \ldots u_k$, where each $u_i$ (except possibly $u_k$) is the longest prefix of $u_iu_{i+1} \ldots u_k$ that has another occurrence to the left in $w$ or a single letter in case this prefix is empty. For our example the Lempel–Ziv factorization is $a.b.b.a.abb.baa.ab.ab$.

The Lempel–Ziv factorization is a basic and powerful technique for text compression [17]. It has many
variants used in gzip or PKZip software, and more generally in dictionary compression methods.

The Lempel–Ziv factorization is easily computed from LPF. The algorithm is shown in Fig. 5. For the example text abbaabbabaaab in Fig. 1, the algorithm outputs $LZ = [0, 1, 2, 3, 4, 7, 10, 12]$.

**Theorem 2.** The Lempel–Ziv factorization of a string of length $n$ over an integer alphabet can be computed in $O(n)$ time.

The experimental results of [2] may be explained by the fact that suffix-tree-based algorithms computing the Lempel–Ziv factorization usually have complexity $O(|A|n)$, where $|A|$ is the cardinality of the alphabet.

Also, note that suffix trees allow online computation of the Lempel–Ziv factorization. However, this comes with the extra $\log(|A|)$ factor. It remains open whether true linear-time online computation is possible.

6. **Application 2: Computing runs in linear time**

Repetitions are a fundamental topic in stringology and appear in many applications such as text algorithms, data compression, or analysis of biological sequences. The simplest repetition is a square $ww$, where $w$ is any string. A general repetition has the form $w^e$, for any rational exponent $e \geq 2$ such that $e|w|$ is an integer; e.g., $(aabab)^{7/5} = aababa$. Particularly important turned out to be maximal repetitions [15] or runs. A run is an occurrence of a repetition that cannot be extended. As an example, the string $aababaabba$ contains the runs $aa$ at positions 0 and 5, $ababa$, and $bb$. Runs allow the encoding of all repetitions in linear space [13].

An element of the Lempel–Ziv factorization carries information already processed by any online algorithms computing repetitions. Therefore it is not surprising that the Lempel–Ziv factorization plays a central part in the algorithm of [13] as well as in all efficient computations of repetitions in strings. Their running time is then $O(n \log(|A|))$. Using a suffix array and the algorithms described in previous sections leads to linear-time computations on integer alphabets. This applies to: testing square freeness of a string [3], computing all leftmost maximal periodicities [15], computing all runs [13], computing of all local periods of a string [6], and computing all primitively-rooted squares occurring in a string [8]; see also [4]. In particular, we get:

**Theorem 3.** The runs of a string of length $n$ over an integer alphabet can be computed in $O(n)$ time.

Acknowledgements

We warmly thank Bill Smyth and Simon Puglisi for interesting discussions related to the present subject during the “String Masters” meeting held in Hamilton in July 2007.

Also, we would like to thank the anonymous referees for comments that helped improving the presentation.

References