A simple proof that a word of length n has at most $2n$ distinct squares

Lucian Ilie*

Department of Computer Science, University of Western Ontario
London, ON N6A 5B7, CANADA
ilie@csd.uwo.ca

Fraenkel and Simpson [2] proved that the number of distinct squares in a word of length n is at most $2n$. Their proof[1] uses a rather intricate combinatorial result of Crochemore and Rytter [1] concerning the lengths of three squares which are prefixes of each other. The same proof is included also in Lothaire's second book [4, p.281-2]. We give here a very short direct proof of this important result. Our proof uses only basic facts from combinatorics on words, see [3, 4].

We give first some notation. Fix an alphabet A; the elements of A are called letters. The set of finite words over A is A^* which is a monoid with concatenation (juxtaposition); its unit element is ε, the empty word. The length of w, that is, the number of letters of w, is denoted $|w|$; $|\varepsilon| = 0$. For $x, y, w \in A^*$, if $w = xy$, then x is called a prefix of w; when $x \neq w$, then x is a proper prefix, denoted $x < w$. For a word w and an integer $n \geq 0$, the nth power of w is defined inductively as $w^0 = \varepsilon$, $w^n = ww^{n-1}$. A word w is called primitive if there is no word x and integer $p \geq 2$ such that $w = x^p$. We shall need two very basic properties of primitive words. First, any word can be written uniquely as an integer power of a primitive word. Second, if w is primitive, then w has exactly two occurrences as factor of ww, namely as a prefix and as a suffix. This property is called synchronization. It is proved immediately by noting that, if w appears somewhere in the "middle" of ww, then w can be written as both xy and yx, for some non-empty words x, y. But then $xy = yx$ implies x and y are powers of the same word and so w is not primitive, a contradiction.

Theorem 1 (Fraenkel and Simpson). *Any word of length n has at most $2n$ distinct squares.*

Proof. We shall count each square at the position where its last occurrence starts (as in [2]). It is enough to prove that no three squares can have the last occurrences starting at the same position. Assume they do and we have $w^2 < v^2 < u^2$. Figure 1 makes the reasoning below easy to follow. We must have $u < w^2$ as otherwise w^2 would appear later. Denote the second occurrence of w in w^2 by w_1, the prefix w of the second v by w_2, and the prefix w of the second u by w_3. Put $v = wx^p$, x primitive, $p \geq 1$. The overlap between w_1 and w_2 gives that $w = x^q x'$ for $q \geq p$, $x = x'x''$, $x' < x$. By synchronization, the overlap

* Research supported in part by NSERC.

[1] The proof seems to be the result of a sequence of improvements, according to the acknowledgements in [2].

between w_2 and w_3, longer than $|x^p|$, is then $x^r x'$, for $p \leq r < q$. Therefore, the remaining suffix of w_3, which is $x'' x^{q-r-1} x'$, and x^p, as suffix of the second v – the two grey rectangles in Figure 1 – begin at the same point. Thus one is a prefix of the other, implying, by synchronization, that $x' = \varepsilon$. Notice that either of the two suffixes can be longer; relevant is that both are of length at least $|x|$. Finally, $w^2 = x^{2q}$ appears again $|x|$ positions later, a contradiction. □

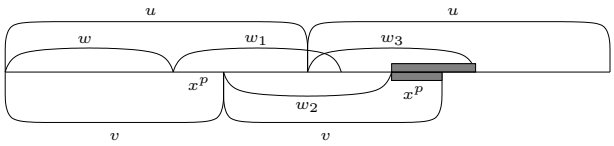

Fig. 1. Three squares at the same position.

Fraenkel and Simpson [2] ask whether the number of distinct squares in a word of length n is in fact at most n, a fact supported also by numerical computations. According to our knowledge, there has been no progress in this direction and any non-trivial improvement of Theorem 1 seems difficult.

References

1. M. Crochemore and W. Rytter, Squares, cubes, and time-space efficient string searching, *Algorithmica* **13** (1995) 405 – 425.
2. A.S. Fraenkel and J. Simpson, How many squares can a string contain?, *J. Combin. Theory, Ser. A,* **82** (1998) 112 – 120.
3. M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, Mass., 1983.
4. M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge Univ. Press, 2002.