1994

A Representation of Approximate Self- Overlapping Word and Its Application

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:
94-053

A REPRESENTATION OF APPROXIMATE SELF-OVERLAPPING WORD AND ITS APPLICATIONS
(Extended Abstract)

Wojciech Szpankowski

Computer Sciences Department
Purdue University
West Lafayette, IN 47907

CSD TR-94-053
August 1994

A Representation of Approximate Self-Overlapping Word and Its Applications

(EXTENDED ABSTRACT)

Wojciech Szpankowski[1]
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
spa@cs.purdue.edu

1. Problem Formulation and Notations

Informally speaking, we are interested in the structure of a word w_k of length k such that when shifted by, say s, the shifted word is within a given distance from the original (un-shifted word). In this note we concentrate on Hamming distance. Later, we deal with the edit distance, too.

We start with some definitions. A word of length, say k, we write as w_k, or more precisely $w_1^k = w_k$. The set of all words of length k is denoted as $\mathcal{W}_k$. Furthermore, a prefix of length $q \le k$ of w_k is denoted as $\overline{w}_k(q)$ or simple $\overline{w}_k$ if there is no confusion.

The distance between words is understood as the relative Hamming distance, that is, $d_n(x_1^n, \widetilde{x}_1^n) = n^{-1}\sum_{i=1}^n d_1(x_i, \widetilde{x}_i)$ where $d_1(x, \widetilde{x}) = 0$ for $x = \widetilde{x}$ and 1 otherwise $(x, \widetilde{x} \in \mathcal{A})$. We also write $M(x_1^n, \widetilde{x}_1^n) = nd_n(x_1^n, \widetilde{x}_1^n)$ for number of mismatches between x_1^n and $\widetilde{x}_1^n$.

Let us now fix $D > 0$. Consider a word $w_{k+s} = w_1^{k+s}$ of length $k + s$, and shift it by $s \le k$. The shifted word of length k is w_s^{k+s}. We would like to identify a set $\mathcal{W}_{k,s}(D)$ of all words w_{k+s} such that

$$d(w_1^k, w_s^{k+s}) \le D \ . \qquad (1)$$

This problem is well understood for "faithful" (lossless) overlapping, that is, when $D = 0$. In this case, we have for $m = \lfloor k/s \rfloor$ (cf. [6, 11, 12])

$$\begin{aligned} \mathcal{W}_{k,s}(0) &= \{w_s \in \mathcal{W}_s : \ w_{k+s} = w_s^{(m+1)}\overline{w}_s \\ &= \bigcup_{w_s \in \mathcal{W}_s} \{w_s^{(m+1)}\overline{w}_s\} \end{aligned} \qquad (2)$$

where $\overline{w}_s$ is a prefix of length $q = k - m \cdot s$, and $w_s^{(m)}$ is a concatenation of m words w_s. Our goal is to extend (2) to the approximate case, that is, for $D > 0$.

[1]This research was supported by NSF Grants NCR-9206315 and CCR-9201078, and in part by NATO Grant 0057/89.

There is plenty of applications of this problem, most notably to approximate pattern matching (cf. [1, 2, 3, 8, 13]) and lossy data compression (cf. [7, 9, 11, 12]). In the former case, Myers [8] observed that to find all approximate pattern matchings of a word w_k (which usually represents a small fraction of the pattern) in a larger text string T, it is enough to generate all words within given distance from w_k and then perform exact pattern matching of every word in such a set and the text string T. We can refine this by considering not only a D-neighborhood of w_k but also a neighborhood of the shifted word, that is, the set $\mathcal{W}_{k,s}(D)$. This refinment is based on a premise that in text T there are regions with approximately repeated structures (e.g., DNA). In order to assess the quality of such an approach, one must estimate the cardinality of $\mathcal{W}_{k,s}(D)$. This is discussed in Section 3.

In a lossy data compression [7] as well as in an approximate pattern matching [2, 3], one is interested in the typical behavior of the longest substring that approximately occurs twice in a given (training or database) sequence. Our representation of the set $\mathcal{W}_{k,s}(D)$ is crucial to establish an upper bound for such a substring. This is discussed in Section 4.

2. Structure of the Word

We construct now all words w_{k+s} that belongs to $\mathcal{W}_{k,s}(D)$. First, let us define an integer ℓ such that $\ell/k \le D < (\ell+1)/k$. Also, we write $k = s \cdot m + q$ where $0 \le q < s$.

Take now $0 \le l \le \ell$, and partition the integer l into $m+1$ integer terms as follows:

$$l = a_1 + a_2 + \cdots + a_m + \tilde{a}_{m+1} \qquad 0 \le a_i \le s \quad \text{for} \quad 1 \le i \le m \tag{3}$$

and $0 \le \tilde{a}_{m+1} \le q$. Clearly, there are many ways of partitioning the integer l into terms as prescribed in (3) (cf. [4]). Let the set of all such partitions be denoted as $\mathcal{P}_{k,s}(l)$.

We now define recursively m sets $\mathcal{W}_s(a_i)$ for $i \le m$. We set $\mathcal{W}_s(a_0) := \mathcal{W}_s$ where $a_0 = 0$. Then,

$$\mathcal{W}_s(a_k) = \{v_s \in \mathcal{W}_s : \ M(w_s, v_s) = a_k \quad \text{for} \quad w_s \in \mathcal{W}_s(a_{k-1})\}\ , \tag{4}$$

and

$$\overline{\mathcal{W}}_q(\tilde{a}_{m+1}) = \{v_q \in \mathcal{W}_q : \ M(\overline{w}_s(q), v_q) = \tilde{a}_{m+1} \quad \text{for} \quad w_s \in \mathcal{W}_s(a_m)\}\ . \tag{5}$$

Now, we can present our main result which follows directly from the above discussion.

Theorem 1.. *Let w_{k+s} be a word such that (1) holds for some $D > 0$. With the notation as above,*

$$\mathcal{W}_{k,s}(D) = \bigcup_{l=0}^{l} \{W_{k,s}(l)\}$$

such that

$$W_{k,s}(l) = \bigcup_{w_s^0 \in \mathcal{W}_s} \bigcup_{\mathcal{P}_{s,k}(l)} \bigcup_{w_s^1 \in \mathcal{W}_s(a_1)1} \cdots \bigcup_{w_s^m \in \mathcal{W}_s(a_m)} \bigcup_{\overline{w}_s^{m+1} \in \overline{\mathcal{W}}_s(\widetilde{a}_{m+1})} w_s^0 w_s^1 \cdots w_s^m \overline{w}_s^{m+1} \qquad (6)$$

where $w_s^0 w_s^1 \cdots w_s^m \overline{w}_s^{m+1}$ *means concatenation of words* w_s^0 *and ... and* $\overline{w}_s^{m+1}$. ■

3. Enumeration

As mentioned in the introduction, to assess complexity of some algorithms dealing with approximate pattern matching one needs to know the cardinality of $\mathcal{W}_{k,s}(D)$. From our Theorem 1 one can easily estimate the cardinality of $\mathcal{W}_{k,s}(l)$ once we know the cardinality the set $\mathcal{P}_{s,k}(l)$.

A. Cardinality of the partition $\mathcal{P}_{s,k}(l)$

The enumeration of $\mathcal{P}_{s,k}(l)$ is not that difficult but rather troublesome. Let $G(z)$ be the generating function of the cardinality $|\mathcal{P}_{s,k}(l)|$ of $\mathcal{P}_{s,k}(l)$. Having in mind the notation as in (3), we immediately obtain the following (cf. [4])

$$\begin{aligned} G(z) &= (1+x+x^2+\cdots+x^s)^m(1+x+x^2+\cdots+x^q) && (7)\\ &= \frac{(1-x^{s+1})^m(1-x^{q+1})}{(1-x)^{m+1}}\,, && (8)\end{aligned}$$

where $m = \lfloor k/s \rfloor$ and $q = k - ms$.

Let $e_l = |\mathcal{P}_{s,k}(l)|$, that is, $e_l = [G(z)]_l$ (coefficient of $G(z)$ at z^l). Following Comtet [4] (cf. Ex. 16 page 77) we introduce *polynomial coefficients* $\binom{n,q}{k}$ as

$$G(x) = (1+x+\cdots+x^{q-1})^n = \sum_{k=0}^{\infty} \binom{n,q}{k} x^k\,. \qquad (9)$$

Note that $\binom{n,2}{k} = \binom{n}{k}$.

Using this and standard generating function arguments we obtain the next lemma.

Lemma 2. *The cardinality* e_l *of* $\mathcal{P}_{s,k}(l)$ *is given by*

$$\begin{aligned} e_l = |\mathcal{P}_{s,k}(l)| &= \sum_{j=0}^{q} \binom{m, s+1}{l-j} && (10)\\ &= \sum_{j=0}^{q} \sum_{(s+1)i+t=l-j} (-1)^i \binom{m}{i}\binom{m+t}{m} && (11)\end{aligned}$$

where $m = \lfloor k/s \rfloor$ *and* $q = k - ms$.

Proof. Formula (10) follows directly from (7) and definition of polynomial coefficients (9). The second enumeration formula (11) is a simple consequence of (8) . ∎

The next interesting question is how to get some asymptotics for e_l. This depends on establishing some asymptotics on the polynomial coefficients. We discuss it in sequel.

We prove the following result. Let $g(z) = (\frac{1}{q} + \frac{z}{q} + \cdots \frac{z^{q-1}}{q})$ be a probability generating function so that the generating function $G(z)$ of polynomial coefficients is $G(z) = q^n g(z)^n$. Clearly, from the Cauchy formula we have

$$\binom{n,q}{k} = \frac{q^n}{2\pi i} \oint \frac{g(z)^n}{z^{k+1}} dz \tag{12}$$

where the path of integration encloses the origin. Judging from the binomial coefficients (i.e., $q = 2$) we should expect different asymptotics for various values of k (e.g., bounded k, k around the mean $n\mu = n(q-1)/2$, and $k = \alpha n$ where $\alpha \neq (q-1)/2$). This is confirmed by the result below.

Lemma 3. *For any q and large n the following holds.*

(i) *If $k = n(q-1)/2 + r$ where $r = o(\sqrt{n})$, then*

$$\binom{n,q}{k} \sim \frac{q^n}{\sigma\sqrt{2\pi n}} \exp\left(-\frac{r^2}{2n\sigma^2}\right) \tag{13}$$

where $\sigma^2 = (q^2-1)/12$. In particular (cf. Comtet [4] [Ex. 16, p.77]),

$$\sup_k \binom{n,q}{k} = \binom{n,q}{n(q-1)/2} \sim q^n \sqrt{\frac{6}{(q^2-1)\pi n}} \,. \tag{14}$$

(ii) *If $k = \alpha n$ where $\alpha \neq (q-1)/2$, then*

$$\binom{n,q}{k} \sim \frac{g(\beta)^n}{\beta^{\alpha n}} \frac{1}{\sigma_\alpha \sqrt{2\pi n}} \tag{15}$$

where β is a solution of $\beta g'(\beta) = \alpha g(\beta)$ and $\sigma_\alpha^2 = \beta^2 g''(\beta)/g(\beta) + \alpha - \alpha^2$.

(iii) *If $k = O(1)$, then*

$$\binom{n,q}{k} \sim \frac{n^k}{k!} \tag{16}$$

Proof. Part (i) is direct consequence of applying the *saddle point method* to the Cauchy integral. Details can be found in Greene and Knuth [5] (page 70-76). Formula (14) comes from the previous one after substitution $r = 0$. Comtet [4] suggests also another derivation of it. Namely, note that after substitution $z = e^{ix}$ and easy algebra the Cauchy formula becomes

$$\binom{n,q}{k} = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2}\left(\frac{\sin(qx)}{\sin(x)}\right)^n \cos(x(n(q-1)-2k))dx \ . \tag{17}$$

Observe that for $k = n(q-1)/2$ the cosine function is equal to one, hence maximum, and then by a simple application of Laplace'a method we get again (14).

Part (ii) follows from (i) and the "method of mean shift" as in Greene and Knuth [5] (page 75). That is, we use part (i) applied to the following

$$[z^{\alpha n}](g(z))^n = \frac{g(\beta)^n}{\beta^{\alpha n}}\left(\frac{g(\beta z)}{g(\beta)}\right)^n$$

where β is a solution of $\beta g_1(\beta) = \alpha g_1(\beta)$.

Part (iii) can be proved as follows. From the Cauchy integral we have after substituting $z/n = w$

$$\begin{aligned}\binom{n,q}{k} &= \frac{1}{2\pi i}\oint \frac{G(z)^n}{z^{k+1}}dz \\ &= \frac{1}{2\pi i}\oint \frac{(1+w/n+\cdots(w/n)^{q-1})^n}{w^{k+1}}n^k dw \to n^k\oint\frac{e^w}{w^{k+1}} = \frac{n^k}{k!} \ .\end{aligned}$$

This completes the proof. ■

Finally, we can formulate our next result that enumerates $\mathcal{W}_{s,k}(l)$.

Theorem 4. *Cardinality of the set $\mathcal{W}_{k,s}(l)$ as defined in (6) is equal to*

$$|\mathcal{W}_{k,s}(l)| = 2^s \sum_{a_1+a_2+\cdots+a_m+a_{m+1}=l}\binom{s}{a_1}\cdots\binom{s}{a_m}\binom{q}{a_{m+1}} = 2^s\binom{k}{l} \ . \tag{18}$$

Proof. The above follows directly from Theorem 1, and the following identity (that we express in generating function terms): $(1+x)^s(1+x)^s\cdots(1+x)^s(1+x)^q = (1+x)^{ms+q} = (1+x)^k$ (cf. [4]). ■

Remark. One can verify our enumeration in Theorem 4. Indeed, we know that summing over all $|\mathcal{W}_{k,s}(l)|$ for $1 \le l \le k$ should give 2^{s+k}, as (18) implies.

4. Typical Behavior of Repeated Patterns

We consider a typical behavior of repeated patterns in an approximate pattern matching (cf. see [7] for applications a lossy data compression, and [2, 3] for applications to approximate pattern matching and DNA sequencing). In particular, we investigate the so called *height* (cf. also [2, 3, 7, 11, 12]). We study the typical behavior of the height in the so called *mixing probabilistic model* as defined in [10, 11, 12] which includes Bernoulli and Markovian models.

More precisely, to define a stationary, ergodic *mixing model* we consider a sequence $\{X_k\}_{k=-\infty}^{\infty}$ that is stationary and ergodic. In addition, it is mixing in strong sense, that is, (informally speaking) for two events A and B defined respectively with σ-algebras of $\{X_k\}_{-\infty}^{m}$ and $\{X_k\}_{m+b}^{\infty}$ for some integer b, the following holds

$$(1-\alpha(b))\Pr\{A\}\Pr\{B\} \leq \Pr\{A\cap B\} \leq (1+\alpha(b))\Pr\{A\}\Pr\{B\}$$

for some some $\alpha(b)$ such that $\lim_{b\to\infty}\alpha(b)=0$.

Let now H_n be the height, that is, the largest K for which there exist $i,j \leq n$ such that $d(X_i^{i+K-1}, X_j^{j+K-1}) \leq D$ where X_1^n is the so called training sequence or "database" sequence that is used in a compression scheme. To express the height in a simple form, we introduce approximate self-overlap C_s as the longest (approximate) prefix of X_1 and X_{1+s} (i.e., a word and its s-shift). More precisely, C_s is the largest K such that $d(X_1^K, X_{1+s}^{K+s}) \leq D$. Observe that C_s is defined with respect to *only* two substrings while H_n with respect to $O(n^2)$ substrings.

In order to estimate the height, we use the following

$$\Pr\{H_n \geq k\} \leq n\left(\sum_{s=1}^{k-1}\Pr\{C_s\geq k\} + \sum_{s=k}^{n}\Pr\{C_s\geq k\}\right) . \tag{19}$$

The second sum is easy to estimate. Indeed,

$$\sum_{s=k}^{n}\Pr\{C_s\geq k\} \leq n\sum_{w_k\in\mathcal{W}_k}P(B_D(w_k))P(w_k) \leq nEP(B_D(w_k)) , \tag{20}$$

where $B_D(w_k)$ is the so called D-ball that contains all words of length k within distance D from the center w_k, that is, $B_D(w_k)=\{x_k :\ d(x_k,w_k)\leq D\}$. By $P(B_D(w_k))$ we denote the probability of the D-ball.

The difficulties arise with the first sum of (19). For this we need a representation of an approximate self-overlapping of a word, which is discussed in sequel (and is of its own interest). In this note we study only an upper bound on H_n (which is a harder part of the analysis). Clearly, $\sum_{s=1}^{k-1}\Pr\{C_s\geq k\} \leq k\Pr\{C_s\geq k\}$ so we need only $\Pr\{C_s\geq k\}$ for $s\leq k$. In this case we have

$$\Pr\{C_s\geq k\} \leq \sum_{w_k\in\mathcal{W}_{k,s}(D)} P(w_k) = \sum_{w_s\in\mathcal{W}_s} P(w_s\widetilde{\mathcal{W}}_{k,s}(D)) \tag{21}$$

where we split the set $\mathcal{W}_{k,s}(D)$ found in our Theorem as $\mathcal{W}_{k,s}(D) = \mathcal{W}_s \cup \widetilde{\mathcal{W}}_{k,s}(D)$.

Now, we proceed as follows

$$\begin{aligned}
\Pr\{C_s \geq k\} &\leq \sum_{w_s \in \mathcal{W}_s} P(w_s \widetilde{\mathcal{W}}_{k,s}(D)) \leq^{(A)} c \sum_{\mathcal{W}_s} P(\widetilde{\mathcal{W}}_{k,s}(D)) P(w_s) \\
&\leq^{(B)} c \sqrt{\sum_{\mathcal{W}_s} P^2(\widetilde{\mathcal{W}}_{k,s}(D)) P(w_s)} \leq c \sqrt{\sum_{\mathcal{W}_s} P(\widetilde{\mathcal{W}}_{k,s}(D)) P(w_s)} \\
&= c\sqrt{EP(\widetilde{\mathcal{W}}_{k,s}(D))} \leq^{(C)} c\sqrt{EP(B_D(w_k))} ,
\end{aligned}$$

where the inequality (A) is due to the mixing condition, inequality (B) is a consequence of the *inequality on means*, and the last inequality (C) follows from $\widetilde{\mathcal{W}}_{k,s}(D) \subset B_D(w_k)$ and hence $EP(\widetilde{\mathcal{W}}_{k,s}(D)) \leq EP(B_D(w_k))$ (in the latter we treat w_s and w_k as random sequences with probability $P(\cdot)$ inherited from the sequence $\{X_k\}$. In the above, the constant c may change from line to line.

In passing, we note that the above estimate can be obtained in a different manner, too. For curiosity, we shall work it out. We start with the second line of the above display to obtain

$$\begin{aligned}
\Pr\{C_s \geq k\} &\leq c\sqrt{\sum_{\mathcal{W}_s} P^2(\widetilde{\mathcal{W}}_{k,s}(D)) P(w_s)} \leq c\sqrt{\sum_{\mathcal{W}_s} P(\widetilde{\mathcal{W}}_{k,s}(D)) P(w_s \widetilde{\mathcal{W}}_{k,s}(D))} \\
&\leq^{(F)} c\sqrt{\sum_{\mathcal{W}_k} P(B_D(w_k)) P(w_k)} = c\sqrt{EP(B_D(w_k))}
\end{aligned}$$

where the inequality (F) follows as before from $\mathcal{W}_s \subset \mathcal{W}_k$, $\widetilde{\mathcal{W}}_{k,s}(D) \subset B_D(w_k)$, and the fact that $w_k = w_s \widetilde{\mathcal{W}}_{k,s}$ for $w_k \in \mathcal{W}_{k,s} \subset \mathcal{W}_k$.

Putting everything together, from the above and (19)-(20), we have

$$\Pr\{H_n \geq k\} \leq nk\sqrt{EP(B_D(w_k))} + n^2 EP(B_D(w_k)) .$$

Therefore, we finally prove that

$$\Pr\{H_n \geq (1+\varepsilon)\frac{2}{r_1(D)} \log n\} \leq \frac{c \log n}{n^{\varepsilon}} \tag{22}$$

where, in general, for any integer $b \neq 0$ we have

$$r_b(D) = \lim_{k\to\infty} \frac{-\log\left(\sum_{w_k \in \mathcal{W}_k} P^b(B_D(w_k))P(w_k)\right)}{bk} = \lim_{k\to\infty} \frac{-\log EP^b(B_D(w_k))}{bk} . \tag{23}$$

The above limit exists due to mixing condition and submultiplicativity of $P(B_D(w_k))$. For $b = 0$ we have from the above by taking $b \to 0$

$$r_0(D) = \lim_{k\to\infty} \frac{-\sum_{w_k \in \mathcal{W}_k} P(w_k) \log P(B_D(w_k))}{k} = \lim_{k\to\infty} \frac{-E \log P(B_D(w_k))}{k} . \tag{24}$$

We can summarize our finding in the following which extends the result of [2] to mixing model.

Theorem 5. *Let X_1^n be a sequence of length n generated according to the mixing probabilistic model. Then, $H_n/\log n \le 2/r_1(D)$ (pr.) where $r_1(D)$ is defined above. In fact, we can proved that $H_n/\log n \to 2/r_1(D)$ (pr.) as $n \to \infty$, and actually the latter limit holds also in almost sure sense.* ∎

References

[1] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, A Basic Local Alignment Search Tool, *J. Molecular Biology*, 215, 403-410, 1990.

[2] R. Arratia and M. Waterman, The Erdös-Rényi Strong Law for Pattern Matching with Given Proportion of Mismatches, *Annals of Probability*, 17, 1152-1169 (1989).

[3] R. Arratia, L. Gordon, and M. Waterman, The Erdös-Rényi Law in Distribution for Coin Tossing and Sequence Matching, *Annals of Statistics*, 18, 539-570 (1990)

[4] L. Comtet, *Advanced Combinatorics*, D. Reidel Publishing Company, Boston, 1974.

[5] D.H. Greene and D.E. Knuth, *Mathematics for the Analysis of Algorithms*, Birkhauser, 1981

[6] M. Lothaire, *Combinatorics on Words*, Addison-Wesley (1982).

[7] T. Luczak and W. Szpankowski, A Lossy data Compression Based on String Matching. Preliminary Analysis and Suboptimal Algorithms, *Proc. Combinatorial Pattern Matching*, Asilomar, California, 1994.

[8] E. Myers, A Sublinear Algorithm for Approximate Keyword Searching, *Algorithmica*, to appear.

[9] D. Ornstein and P. Shields, Universal Almost Sure Data Compression, *Annals of Probability*, 18, 441-452 (1990).

[10] B. Pittel, Asymptotic Growth of a Class of random Trees, *Annals of Probability*, 13, 414 - 427 (1985).

[11] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, 39, 1647-1659 (1993).

[12] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176-1198 (1993).

[13] E. Ukkonen, Approximate String-Matching over Suffix Trees, *Proc. Combinatorial Pattern Matching Conference*, Padova, LNCS 684, 228-242, Springer-Verlag 1993.