

New and Old Problems in Pattern Matching

Wojciech Szpankowski

Computer Science Department, Purdue University (USA)

June 25, 2001

Summary by Mireille Régnier

Abstract

This talk presents three problems in pattern matching and their analysis. Different methods are used, that rely on complex analysis and probability theory.

1. Statement of the Problems

Some pattern H (or a set \mathcal{H} of patterns) is searched in a text T . The text T is generated by a random probabilistic source that is either a Bernoulli source or a Markov source or a mixing source. In the string matching and the subsequence matching problems, H is given: the model is deterministic. In the *repetitive patterns problem*, in Section 4, H is a string of T repeated elsewhere.

2. String Matching

One counts the number of occurrences of a given word H or a given finite set of words, \mathcal{H} , in a text of size n . This number is denoted $O_n(H)$ or $O_n(\mathcal{H})$. This counting relies on the decomposition of the text T onto languages, the so-called *initial*, *minimal*, and *tail languages*.

Definition 1. Given two strings H and F , the *overlap set* is the set of suffixes of H that are also prefixes of F . The suffixes of F in the associated factorizations of F form the *correlation set* $\mathcal{A}_{H,F}$. In the Bernoulli model, one defines the *correlation polynomial* of H and F as

$$A_{H,F}(z) = \sum_{w \in \mathcal{A}_{H,F}} P(w)z^{|w|}.$$

When H is equal to F , $\mathcal{A}_{H,H}$ is named the *autocorrelation set* and denoted $\mathcal{A}_{H,H}$; the *autocorrelation polynomial* is defined as

$$A_H(z) = \sum_{w \in \mathcal{A}_{H,H}} P(w)z^{|w|}.$$

For example, let $H = 11011$ and $F = 1110$. Then the overlap set of H and F is $\{11, 1\}$ and the correlation set is $\mathcal{A}_{H,F} = \{10, 110\}$. Similarly, $\mathcal{A}_{F,H} = \{11\}$. It is worth noticing that $\mathcal{A}_{F,H} \neq \mathcal{A}_{H,F}$. Intuitively, the concatenation of a word in $\mathcal{A}_{H,F}$ to H creates an (overlapping) occurrence of F .

Definition 2. Let H be a given word.

- (i) The *initial* language \mathcal{R} is the set of words containing only one occurrence of H , located at the right end.

- (ii) The *tail language* \mathcal{U} is defined as the set of words u such that Hu has exactly one occurrence of H , which occurs at the left end.
- (iii) The *minimal language* \mathcal{M} is the set of words w such that Hw has exactly two occurrences of H , located at its left and right ends.

With these notations, any text that contains exactly k occurrences of H , $k \geq 1$, rewrites unambiguously as

$$rm_1 \dots m_{k-1}u$$

where $r \in \mathcal{R}$, $m_i \in \mathcal{M}$, and $u \in \mathcal{U}$. In other words, this set \mathcal{T}_k of words satisfies $\mathcal{T}_k = \mathcal{R}\mathcal{M}^{k-1}\mathcal{U}$. The power of this approach comes from the equations that can be written on these languages, that translate into equations on their generating functions in the Bernoulli model *and* the Markov model. Moreover, it turns out that these generating functions—hence the whole counting problem—only depend on the probability of H , denoted $P(H)$, and the so-called correlation set.

Theorem 1. *Let H be a given pattern of size m , and T be a random text generated by a Bernoulli model. The generating function of the set \mathcal{T}_k satisfies*

$$T_k(z) = z^m P(H) \frac{(D_H(z) + 1 - z)^{k-1}}{D_H(z)^{k+1}}, \quad k \geq 1,$$

$$T_0(z) = \frac{A_H(z)}{D_H(z)}$$

where

$$D_H(z) = (1 - z)A_H(z) + z^m P(H).$$

Moreover, the bivariate generating function satisfies

$$T(z, u) = \sum_k T_k(z) u^k = \frac{u}{1 - u \frac{D_H(z) + 1 - z}{D_H(z)}} \frac{z^m P(H)}{D_H(z)^2}$$

These results extend to the Markovian model and to the case of multiple pattern matching [3].

3. Subsequence Matching

A pattern $W = w_1 \dots w_m$ is hidden in a text T if there exist indices $1 \leq i_1 < \dots < i_m \leq n$ such that $t_{i_1} = w_1, \dots, t_{i_m} = w_m$. For example, *date* is hidden 4 times in the text *hidden pattern* but it is not a substring. We focus on cases where the sequence of indices satisfies additional constraints $i_{j+1} - i_j \leq d_j$, where d_j is either an integer or ∞ . Such a sequence is called an *occurrence*. One denotes (d_1, \dots, d_{m-1}) by \mathcal{D} . For example, when $\mathcal{D} = (3, 2, \infty, 1, \infty, \infty, 4, \infty)$ the set $I = (5, 7, 9, 18, 19, 22, 30, 33, 50)$, satisfies the constraints.

The number of occurrences, Ω_n , is asymptotically Gaussian. This is proved in [1] by the moments method: all moments of the properly normalized random variable converge to the corresponding moments of the Gaussian law. For any sequence I that satisfies the constraints, one denotes X_I the random variable that is 1 if $t_{i_1} = w_1, \dots, t_{i_m} = w_m$. Then,

$$\Omega_n = \sum_I X_I.$$

The computation of the moments relies on a generalization of correlation sets. Let

$$\mathcal{U} = \{u_1, \dots, u_{b-1}\}$$

be the subset of indices j for which $d_j = \infty$. Any occurrence I satisfying the constraints can be divided into b blocks:

$$[i_1, i_{u_1}], [i_{u_1+1}, i_{u_2}], \dots, [i_{u_{b-1}+1}, i_m].$$

The collection of these blocks is called the *aggregate* of I and denoted $\alpha(I)$. In the example above, the aggregate $\alpha(I)$ is

$$\alpha(I) = [5, 9], [18, 19], [22], [30, 33], [50].$$

Deriving the mean. The collection of occurrences of W can be described as

$$\mathcal{A}^* \times \{w_1\} \times \mathcal{A}^{\leq d_1} \times \{w_2\} \times \dots \times \mathcal{A}^{\leq d_{m-1}} \times \{w_m\} \times \mathcal{A}^*,$$

where \mathcal{A} is the alphabet and $\mathcal{A}^{\leq d_j}$ is the collection of words of size less than or equal to d_j . It follows that the generating function of expectations is

$$\sum_n \mathbf{E}(\Omega_n) z^n = \frac{1}{(1-z)^{b-1}} \times \prod_{i=1}^m p_{w_i} z \times \prod_{i \notin \mathcal{U}} \frac{1-z^{d_i}}{1-z},$$

where p_{w-i} is the probability of character w_i . Hence, the expectation satisfies

$$(1) \quad \mathbf{E}(\Omega_n) = \frac{n^b}{b!} \prod_{i \notin \mathcal{U}} d_i \prod_{i=1}^m p_{w_i} \left(1 + O\left(\frac{1}{n}\right) \right)$$

Deriving the variance and higher moments. The variance rewrites

$$\mathbf{Var}(\Omega_n) = \sum_{I, J} \mathbf{E}(X_I X_J) - \mathbf{E}(X_I) \mathbf{E}(X_J).$$

In the Bernoulli model, the two random variables X_I and X_J are independent whenever the blocks of I and J do not overlap. Hence, the contribution to the variance is zero. If $\alpha(I)$ and $\alpha(J)$ overlap, one defines the aggregate $\alpha(I, J)$ as the set of blocks obtained by merging the blocks of $\alpha(I)$ and $\alpha(J)$ that overlap. The number of blocks in $\alpha(I, J)$, denoted $\beta(I, J)$, is upper bounded by $2b - 1$. For such a pair (I, J) , the text can be rewritten as an element of the language

$$\mathcal{A}^* \times \mathcal{B}_1 \times \mathcal{A}^* \times \dots \times \mathcal{B}_{\beta(I, J)} \times \mathcal{A}^*$$

and the generating function of the covariance rewrites

$$\sum_n \mathbf{Var}(\Omega_n) z^n = \sum_{p \geq 1} \sum_{\beta(I, J) = 2b-p} \frac{1}{(1-z)^{2b-p}} P_p(z),$$

where P_p are polynomials of the variable z that generalize the correlation polynomials defined in [2] (see Definition 1). The asymptotic order of each term is n^{2b-p} . Hence, the dominating contribution is due to the intersecting pairs such that $\beta(I, J) = 2b - 1$, and

$$\mathbf{Var}(\Omega_n) \sim n^{2b-1} \sigma^2$$

where the variance coefficient σ can be easily evaluated for any given pattern by dynamic programming.

The proof is similar for higher moments.

4. Repetitive Pattern Matching

Given a pattern H found in a text T , one searches for a second *approximate* occurrence of H . A word F is a D -approximate occurrence of a word H if the Hamming distance between F and H is smaller than D . Recall that the Hamming distance between two words of size m , say $H = H_1 \dots H_m$ and $F = F_1 \dots F_m$ is

$$d_H(H, F) = \sum_{i=1}^m 1_{H_i \neq F_i}.$$

The usual parameters on trees, such as the *depth of insertion*, *height*, *fill-up*, \dots , are extended in the approximate case. Notably:

Definition 3. The *depth* L_n is the largest integer K such that

$$\min \left\{ d(T_i^{i-K+1}, T_n^{n+K}) \mid 1 \leq i \leq n - K + 1 \right\} \leq D.$$

Rényi's entropy is generalized. Given a word H , the D -ball with center H , denoted $B_D(H)$, is the set of words that are within distance D .

Definition 4. Given a text T , Rényi's entropy of order 0 is

$$r_0(D) = \lim_{k \rightarrow \infty} \frac{-\mathbf{E} \left[\log \mathbf{P} (B_D(T_1^k)) \right]}{k},$$

when this limit exists.

Asymptotic properties are proved for the depth, the height and the fill-up, that depend on Rényi's entropy. Notably, the convergence in probability of the depth of insertion in a trie extends for this approximate scheme:

$$\frac{L_n}{\log n} \rightarrow \frac{1}{r_0(D)}, \quad n \rightarrow \infty.$$

The proof relies on the subadditive ergodic theorem and asymptotic equipartition property.

Bibliography

- [1] Flajolet (P.), Guivarc'h (Y.), Vallée (V.), and Szpankowski (W.). – Hidden patterns statistics. In *ICALP'01*. – 2001. Proceedings of the 28th ICALP Conference, Crete, Greece, July 2001.
- [2] Guibas (L. J.) and Odlyzko (A. M.). – String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A*, vol. 30, n° 2, 1981, pp. 183–208.
- [3] Régnier (M.) and Szpankowski (W.). – On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, vol. 22, n° 4, 1998, pp. 631–649. – Average-case analysis of algorithms.