PROBABILISTIC MODELING OF DATA STRUCTURES ON WORDS:
A Reply to Professor Andersson's Letter

Peter Kirschenhofer
Dept. of Algebra&Discrete Math.
Technical University of Vienna
A-1040 Vienna
Austria

Helmut Prodinger
Dept. of Algebra&Discrete Math.
Technical University of Vienna
A-1040 Vienna
Austria

Wojciech Szpankowski*
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

Professor Arne Andersson's Letter-to-the-Editor concerning our paper "On the Balance Property of Patricia Tries: External Path Length Viewpoint" *Theor. Comp. Sci.*, 68, 1989 motivated us to present some thoughts about probabilistic analysis of data structures on words. The intention of this note is to discuss potential advantages and disadvantages of probabilistic analyses, and in particular to provide a proper interpretation of probabilistic results. This can only be achieved after building a *suitable* probabilistic model for a *given* set of data. We describe a sequence of probabilistic models with an increasing generality that can be applied to the analysis of algorithms on words. Finally, we discuss a few theoretical results to indicate that our findings from the above cited paper hold true under more general probabilistic assumptions.

1. PROBABILISTIC MODELS

It is important to realize that the expected size and shape of the data structures on words depends upon the various probabilistic assumptions on the words (strings). Primarily, there are three considerations:

- Characteristics of the alphabet; that is, the distribution on the alphabet Σ that determines how the symbols are selected to form a string;
- Statistical dependency between words; that is, whether or not the words are statistically dependent;
- Number of words; that is, whether the number n of words is fixed or a random variable.

*This research was supported by the NSF Grant INT-8912631, and in part by AFOSR grant 90-0107, by the NSF grant CCR-9201078 and NCR-9206315, and by grant R01 LM05118 from the National Library of Medicine.

We now discuss these considerations and present a few basic probabilistic models for data structures on words, in particular tries and suffix trees (cf. [1], [2]). To recall, a trie is a digital tree built from a set of keys (strings, sequences) $X_1, \ldots, X_n$. Every key is a sequence of symbols from an alphabet. A suffix tree is a trie built from suffixes of a *single* string (key, sequence) X. To simplify our further presentation, we write X or $\{X_k\}_{k=1}^{\infty}$ as a generic notation for a sequence of symbols from a given alphabet Σ.

The *Bernoulli model* is characterized by the following probabilistic assumptions:

(A) BERNOULLI MODEL

The symbols of the alphabet occur independently of one another; thus, a key $X = x_1 x_2 x_3 \ldots$ can be described as the outcome of an infinite sequence of Bernoulli trials in which $\Pr\{x_j = \omega_i\} = p_i$ and $\sum_{i=1}^{V} p_i = 1$, where $\omega_i \in \Sigma$ and V is the size of Σ. If $p_1 = p_2 = \ldots = p_V = 1/V$, then the model is called *symmetric*; otherwise, it is *asymmetric*.

(B) The number of strings (keys) is fixed at n.

If the following assumption is added to (A) and (B), then such a model is called *independent* Bernoulli model, and it is often used to analyze independent tries.

(C) INDEPENDENT MODEL

The words are statistically independent; that is, knowing a particular string does not give information about the other strings.

In the analysis of tries, assumption (B) leads to quite complicated functional and/or differential-functional equations (cf. [4], [5], [6], [7], [8], [9], [12], [13], [14], [17]). They can be drastically simplified if (B) is replaced by the following assumption.

(B1) POISSON MODEL

The number of strings is a random variable N that has a Poisson distribution.

The Bernoulli model is a good model when keys are transformed into a random stream of bits through a hashing function. However, in many other applications, assumption (A) is not very realistic. For instance, if the strings are words from the English language, then there certainly is a dependence among the symbols of the alphabet. As an example, h is much more likely to follow an s than a b. When this is the case, assumption (A) can be replaced by

(A1) Markovian Model

There is a Markovian dependency between consecutive symbols in a key; that is, the probability $p_{ij} = \Pr\{X_{k+1} = \omega_j | X_k = \omega_i\}$ describes the conditional probability of sampling symbol $\omega_j \in \Sigma$ immediately after symbol $\omega_i \in \Sigma$.

Note that all the models discussed so far have independent strings and are useful for studying tries and PATRICIA tries but not the suffix and PAT trees (i.e., a compact suffix tree; cf. [7]). In the case of suffixes, the strings involved are substrings of a given word so they are clearly dependent. To model this situation, we replace assumption (C) by the following one.

(C1) Dependent Model

The strings are dependent; i.e., knowledge of one key gives information about other strings.

There are two further generalizations of the Markovian model, namely the *mixing model* and the *stationary model*. These models are defined as follows.

(A2) Mixing Model

The sequence $\{X_k\}_{k=1}^{\infty}$ satisfies the so called mixing condition, that is, there exist two constants $0 < c_1 \leq c_2$ and an integer d such that for all $1 \leq m \leq m+d \leq n$ the following holds

$$c_1 \Pr\{\mathcal{A}\}\Pr\{\mathcal{B}\} \leq \Pr\{\mathcal{A}\mathcal{B}\} \leq c_2 \Pr\{\mathcal{A}\}\Pr\{\mathcal{B}\} \tag{1}$$

with $\mathcal{A} \in \mathcal{F}_1^m$ and $\mathcal{B} \in \mathcal{F}_{m+d}^{\infty}$ where $\mathcal{F}_m^n$ is a σ-field generated by $\{X_k\}_{k=m}^n$ for $m \leq n$.

In some statements of results one may need a stronger form of the above mixing condition which is defined in sequel.

(A3) Strong Mixing Model

The sequence $\{X_k\}_{k=1}^{\infty}$ satisfies the so called *strong α-mixing condition* if (1) is replaced by

$$(1-\alpha(d))\Pr\{\mathcal{A}\}\Pr\{\mathcal{B}\} \leq \Pr\{\mathcal{A}\mathcal{B}\} \leq (1+\alpha(d))\Pr\{\mathcal{A}\}\Pr\{\mathcal{B}\} \tag{2}$$

where the function $\alpha(d)$ is such that $\alpha(d) \to 0$ as $d \to \infty$.

In words, model (A2) says that the dependency between two substrings of a given key, say $\{X_k\}_{k=1}^m$ and $\{X_k\}_{k=m+d}^{\infty}$, is rather weak (note that when the sequence $\{X_k\}$ is i.i.d., then $\Pr\{\mathcal{AB}\} = \Pr\{\mathcal{A}\}\Pr\{\mathcal{B}\}$). Assumption (A3) says that this dependency is weaker and weaker as d becomes larger. The "quantity" of dependency is characterized by $\alpha(d)$.

The most general probabilistic model that can provide some useful results, is the stationary model. This model is widely used in coding and communications, most notably in data compression (cf. [18], [21]).

(A4) Stationary Model

The sequence $\{X_k\}_{k=1}^{\infty}$ of letters from a finite alphabet is a *stationary and ergodic* sequence of random variables.

To explain how the stationary model works, we need to introduce some notations. Let $X_m^n = (X_m, \ldots, X_n)$ for $m < n$, and let for every $n \geq 1$ the nth order probability distribution for $\{X_k\}$ be

$$P(X_1^n) = \Pr\{X_k = x_k,\ 1 \leq k \leq n,\ x_k \in \mathcal{A}\}\ .$$

In the stationary model, this probability does not depend on time-shift, that is, if τ is an integer, then for every n and τ the following holds

$$P(X_{1+\tau}^{n+\tau}) = P(X_1^n)\ .$$

For PATRICIA tries and compact suffix trees (i.e., PAT trees) another assumption which strengthens (A3) is often useful.

(P) Contractive Mixing Model

Let $\omega_i \in \Sigma$ for $1 \leq i \leq n$, and define $P(\omega_1, \ldots, \omega_n) = \Pr\{X_1 = \omega_1, \ldots, X_n = \omega_n\}$. Then, for PAT trees we shall require the following condition

$$P(\omega_1, \ldots, \omega_n) \leq \rho P(\omega_1, \ldots, \omega_{n-1}) \tag{3}$$

for some $0 < \rho < 1$.

Finally, the most general probabilistic model makes no assumption regarding the probabilistic nature of keys $X_1, \ldots, X_n$. Then, however, a chance for any interesting results is very dim. Does this model reflect practical situations? We doubt.

In passing, we should mention one more model, which might be called the *worst-case model*. In this case, we assume that for a *given* algorithm (not necessary a problem!) the whole probability mass is concentrated on one (pathological) input that returns the maximum value of the parameter under consideration, say the external path length. Note, however, that such a distribution depends on the algorithm (and the problem) itself.

2. SOME RESULTS

In the discussed paper [11] we have established the asymptotic expansion for the variance of the external path length in a PATRICIA tree (see [10] for similar results for tries) in the *symmetric* Bernoulli model. So far, an extension of this result to an asymmetric Bernoulli model have failed, however, Jacquet and Régnier [8] proved that the limiting distribution for the external path length is normal with the variance $O(n \log n)$ (but the constant hidden in $O(\cdot)$ was not found in [8]). As a simple consequence of our result in [11], we showed that the external path length L_n almost surely behaves like $n \log_2 n$; more precisely $L_n / n \log_2 n \to 1$ (a.s.) as $n \to \infty$. This result is weaker than our main finding in [11], and in fact it holds in a much more general probabilistic setting.

Indeed, Shields [16] proved in the Markovian model that the external path length in a *suffix tree* (hence, also in tries) asymptotically converges to $(1/h) n \log n$ (a.s.) where h is the entropy of the alphabet (precisely: $L_n / n \log n \to 1/h$ (a.s.) as $n \to \infty$). This is also true for PAT trees and PATRICIA (cf. [19]). (Note that $h = \log 2$ for the binary symmetric model.) However, Shields [16] showed also that the above result does *not* hold in general (e.g., in stationary model).

Probably, for independent tries (and PATRICIA) Shields' finding can be extended to the mixing model (but not to the general stationary model as already noted by Shields [16]). Pittel proved in [15] that in the mixing model the typical depth D_n (i.e., the length of a path from the root to a randomly selected external node) behaves asymptotically as $D_n / \log n \to 1/h$ in probability (pr.). For the Markovian model (more precisely: the strong mixing model with the coefficient of mixing $\alpha(d)$ from (A3) decaying exponentially fast) Szpankowski [19] extended Pittel's result to suffix trees (and PAT trees). Since the external path length is the sum of depths, one should expect that it converges (a.s.) to $(1/h) n \log n$ even in the mixing model. (Some caution is adviced in extending the convergence in probability results of D_n to almost sure convergence of the external path length. In fact, as observed by Pittel [15] and Szpankowski [18] (cf. [19] and [21]) $D_n / \log n$ does not converge (a.s.) to $1/h$ but rather oscillates between two constants. Nevertheless, the sum of depths, that is, the external path, should converge (a.s) to $1/h$.) In summary, the conclusions in our paper [11] were restricted to the Bernoulli symmetric model, even if it was later assessed that they hold under weaker probabilistic assumptions.

In passing, we should mention that probabilistic modeling provides good insights into the behavior of algorithms and data structures only when applied to a *set* of inputs, not a *particular* instant of input.

References

[1] A.V. Aho, J.E. Hopcroft and J.D. Ullman, *The Design and Analysis of Computer Algorithms* , Addison-Wesley (1974).

[2] A. Apostolico, The Myriad Virtues of Suffix Trees, *Combinatorial Algorithms on Words*, 85-96, Springer-Verlag, ASI F12 (1985).

[3] W. Chang, E. Lawler, Approximate String Matching in Sublinear Expected Time, *Proc. of 1990 FOCS*, 116-124 (1990).

[4] P. Flajolet, On the Performance Evaluation of extendable Hashing and Trie Searching, *Acta Informatica*, 20, 345-369 (1983).

[5] P. Flajolet and B. Richmond, Generalized Digital Trees and Their Difference-Differential Equations, *Random Structures and Algorithms*, 3, 1992, in press.

[6] P. Flajolet and R. Sedgewick, Digital Search Trees Revisited, *SIAM J. Computing*, 15, 748-767 (1986).

[7] G.H. Gonnet and R. Baeza-Yates, *Handbook of Algorithms and Data Structures*, Addison-Wesly, Workingham (1991).

[8] P. Jacquet and M. Régnier, Normal Limiting Distribution for the Size and the External Path Length of Tries, INRIA TR-827, (1988).

[9] P. Kirschenhofer and H. Prodinger, Further Results on Digital Search Trees, *Theoretical Computer Science*, 58, pp. 143-154 (1988).

[10] P. Kirschenhofer, H. Prodinger and W. Szpankowski, On the Variance of the External Path Length in a Symmetric Digital Trie, *Discrete Applied Mathematics*, 25, 129-143, 1989.

[11] P. Kirschenhofer, H. Prodinger and W. Szpankowski, On the Balance Property of Patricia Tries: External Path Length Viewpoint, *Theoretical Computer Science*, 68, 1-17, 1989

[12] D. Knuth, *The Art of Computer Programming. Searching. Fundamental Algorithm* , vol. I, Addison-Wesley, Reading (1967).

[13] D. Knuth, *The Art of Computer Programming. Sorting and Searching* , vol. III, Addison-Wesley (1973).

[14] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, (1992).

[15] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Annals of Probability* , 18, 414 - 427 (1985).

[16] P. Shields, Entropy and Prefixes, *Annals of Probability*, 20, 403-409 (1992).

[17] W. Szpankowski, Patricia tries again revisited, *Journal of the ACM*, 37, 691-711 (1991).

[18] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, submitted.

[19] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, to appear.

[20] M. Waterman, *Mathematical Methods for DNA Sequences*, CRC Press Inc., Boca Raton, (1991).

[21] A. Wyner and J. Ziv, Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression, *IEEE Trans. Information Theory*, 35, 1250-1258 (1989).