DIGITAL DATA STRUCTURES AND ORDER STATISTICS

Wojciech Szpankowski*
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA

Abstract

This paper studies in a probabilistic framework some topics concerning the way words (strings) can overlap, and relationship of it to the height of digital trees associated with this set of words. A word is defined as a random sequence of (possible infinite) symbols over a finite alphabet. A key notion of alignment matrix \( \{C_{ij}\}_{i,j=1}^{n} \) is introduced where \( C_{ij} \) is the length of the longest string that is prefix of the \( i \)-th and the \( j \)-th word. It is proved that the height of an associated digital tree is simply related to the alignment matrix through some order statistics. In particular, using this observation and proving some inequalities for order statistics, we establish that the height of a digital trie under independent model (i.e., all words are statistically independent), is asymptotically equal to \( 2 \log_{\alpha} n \) where \( n \) is the number of words stored in the trie and \( \alpha \) is a parameter of the probabilistic model. Some extensions of our basic model to other digital trees such as \( b \)-tries, tries with random number of keys (Poisson model) and suffix trees (dependent keys !) are also shortly discussed.

1. INTRODUCTION

Correlation on words are usually studied through the associated data structures namely digital trees built over these words, such as radix tries, subword trees, suffix trees, etc. [1,2,3]. Digital trees are important by their own right due to many applications in computer science (e.g., searching and sorting [1,2], dynamic hashing [4,5], pattern matching algorithms [1,3], etc.) and telecommunications (e.g., coding, conflict resolution algorithms for broadcast communications [6,7,8], etc.). We shall argue that the most interesting parameter, from the complexity viewpoint, is the height of digital trees. In this paper, we show that the height is simply related to the longest common prefix of any two words stored in the tree, and we shall explore this relationship. The key notion of alignment matrix \( C = \{C_{ij}\}_{i,j=1}^{n} \) is introduced, where \( n \) is the number of words (keys, strings) and \( C_{ij} \) measures the overlap on the first symbols in the \( i \)-th and the \( j \)-th words. We shall study properties of the alignment \( C_{ij} \) in a probabilistic framework, that is, it is assumed that words (keys) form a random sequence of (possible infinite) symbols over a finite alphabet. The symbols occur independently in a word, however, words might be statistically dependent (for details see Section 2).

By proving some theorems on order statistics (i.e., maximum) of dependent random variables (i.e., alignments \( C_{ij} \)), we shall establish in this paper a new methodology to study the height of digital trees and some other related problems (e.g., the longest prefix of any pair of words, the longest substring that can be fully recopied, testing for square-free words, memory requirements in the extendible

* This research was supported in part by NSF grants NCR-8846388 and CCR-8900305.
In particular, we prove that for large $n$, the height $H_n$ of a digital trie with independent keys is almost surely equal to $2 \log_2 n$ where $\alpha$ is a parameter of the probabilistic model. In our full paper [15], this result is generalized into three directions, namely in addition we study there $b$-tries, tries with random number of keys (Poisson model) and suffix trees (see also [16]). For example, in [15] and [16] we show that for a suffix tree (in this case keys are statistically dependent) almost surely the height $H_n$ does not exceed $2 \log_2 n$ for some $\kappa$ (see also Section 3).

The height of digital trees has been recently investigated in [2, 5, 9-15]. In [5] Flajolet studied binary symmetric $b$-tries. Based on some classical counting results in occupancy problems, Flajolet derived asymptotic distribution of the height. Using complex analysis (Cauchy integral formula) he also found the average height of a trie. Jacquet and Regnier [9], extended Flajolet's result to binary asymmetric (i.e., symbols occur with different probabilities) tries. They have made extensive use of the Mellin transform technique. Devroye [10] analyzed binary symmetric tries, and based on the occupancy problem he derived some inequalities on the asymptotic distribution of the height. The most general results were obtained by Pittel [11] (see also [12]), where general asymmetric tries with $b = 1$ were investigated. Unfortunately, the proofs in [11] and [12] are not constructive and the results are well hidden. For some more results, see also [13] and [14]. Our approach to compute the height of digital trees is quite different in comparison with the ones established in [2, 9-14]. In contrast to the previous analyses, we use here some novel results from order statistics, and therefore we avoid explicit computation of the height distribution. Finally, we point out that all results discussed so far have been established for independent models, that is, for statistically independent keys. To the best of our knowledge, the dependent models were only studied by Szpankowski [15], and Apostolico and Szpankowski [16]. In [16] the authors investigate the height of suffix trees using the technique established in this paper.

2. MODEL FORMULATION

In this section we build our probabilistic framework, which sets up a stochastic model for our studies. Let $\mathcal{A} = \{\omega_1, \omega_2, \ldots, \omega_V\}$ be an alphabet of $V$ symbols, and let $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ be a set of $n$ (possibly infinite) strings (keys, words, sequences) over the alphabet $\mathcal{A}$. In our basic probabilistic model we assume:

(i) A word $X_k = x_1 x_2 x_3 \cdots$, is an infinite sequence of symbols from $\mathcal{A}$ such that it forms an independent sequence of Bernoulli trials with probability of sampling symbol $\omega_i$ equal to $p_i$, where $\sum_{i=1}^{V} p_i = 1$, that is, $p_i = \Pr(x_i = \omega_i)$ for any $k$ and $j$. If $p_1 = p_2 = \cdots = p_V = 1/V$, then the model is called symmetric, otherwise it is asymmetric.

(ii) The words $X_1, X_2, \ldots, X_n$ are statistically independent.

(iii) The number of words is fixed and equal to $n$.

These three assumptions form our basic probabilistic framework called Bernoulli model. Some modifications of this basic model might be considered. For example, one can replace (iii) by a more general assumption.