On Tree-Growing Search Strategies

Hosam M. Mahmoud
Algorithms Project, INRIA Rocquencourt and The George Washington University

January 5, 1998

[summary by Michèle Soria]

Abstract

A search algorithm that adds a key to a sorted file can be represented by a *deterministic* tree whose external nodes are equally likely targets for insertion. The collection of algorithms that one uses throughout the stages of insertion sort is called a *search strategy*. Using the concept of "tree-growing" strategy, we demonstrate that most practical algorithm have a normal behavior. We present a sufficient condition for normality of tree-growing strategies. The sufficient condition specifies a relationship between the overall variance and the rate of growth in height of the sequence of trees that the search strategy "grows".

Insertion sort is a well-known on-line sorting algorithm: at each stage of the sorting, the elements obtained so far make up a sorted array; when reading a new element, the algorithm searches for its proper position in the array and inserts it. The searching may be done by any method (linear, binary, etc) and the methods may be different from one stage to another. At each stage, given a searching strategy, the positions of *probes* (positions for comparisons between the new element to be inserted and elements of the current array) is represented by a binary decision tree. The first probe is the root of the decision tree, the two positions of the second probe, at most one on each side of the first probe, become the children of the root and all internal nodes are constructed so forth. The leaves of the tree correspond to the places where new elements are to be inserted. The root-to-leaf paths of the tree thus represent the possible probe sequences of the searching algorithm.

1. Tree-growing and normal strategies

Let S_i denote the searching algorithm at stage i, with corresponding decision tree T_i; the collection of searching algorithms $\mathcal{S} = \{S_i\}_{i=1}^{\infty}$, or equivalently the collection of corresponding decision trees $\mathcal{T} = \{T_i\}_{i=1}^{\infty}$, will be called a *search strategy*. A search strategy is *tree-growing* if, for each positive integer i, the shape of T_{i+1} is obtained by replacing a leaf of T_i by an internal node (with two hanging leaves).

We analyse tree-growing search strategies under the assumption of uniform distribution of the leaves at each stage. Let the random variable C_n be the total number of times $\mathcal{S}$ compares a new element to a probe during the sorting of the first n elements. The class of *normal search strategies* is composed of the strategies for which C_n, once normalized, converges in distribution to the normal law, i.e.

$$\frac{C_n - \mathrm{E}[C_n]}{\sqrt{\mathrm{Var}[C_n]}} \to_{\mathcal{D}} \mathcal{N}(0,1).$$

The following lemma gives a sufficient condition for a tree-growing strategy to be normal. This condition relates the variance of C_n to the height of the decision trees of the search strategy. We denote by h_n the height of tree T_n, and by s_n^2 the variance of C_n.

Lemma 1. *If $h_n = o(s_n)$ then $\mathcal{S}$ is a normal strategy.*

Proof. C_n is the sum of n random variables $(X_i)_{i=1,\ldots,n}$, where X_i denotes the number of comparisons made by S_i. Since each insertion is performed independently of all others, we assume the X_i's to be independent random variables. The proof of Lemma 1 is a technical verification of Lindeberg's condition, which ensures normality:

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n-1} \int_{|X_i|>\epsilon s_n} X_i^2 \, dF_{X_i} = 0.$$

□

Practically for a given strategy, the difficulty lies in computing the variance of C_n, which is the sum of the variances of the X_i's. The most commonly used strategies, linear search and binary search, satisfy the condition of Lemma 1. When linear search is used at every stage, it is easy to show that C_n has average value asymptotic to $n^2/4$, and variance asymptotic to $n^3/36$, whereas h_n equals $n-1$. For binary repeated search strategies, one can easily show that h_n is asymptotic to $\log n$ and the average value of C_n is equivalent to $n \log_2 n$, but the computation of the variance is more intricate and finally leads to $s_n^2 = nA(n) + \mathcal{O}(\log n)$, where $A(n)$ is an oscillating function of bounded magnitude.

There exists tree-growing search strategies which are not normal. In [2], the authors exhibit a strategy that does not satisfy the condition of Lemma 1, and can be shown to be non normal by Feller-Lindeberg condition (see [1, vol. 2, §XV.6]). To ensure normality, some further conditions, which are presented in the next section, are required on the decision trees of the search strategy.

2. Normality of consistent strategies

This section identifies a subclass of tree-growing strategies, the consistent strategies, which are proved to be normal.

Let $\mathcal{T} = \{T_i\}_{i=1}^{\infty}$ be the collection of decision trees corresponding to a search strategy $\mathcal{S}$. For each T_i, we denote by T_{L_i} its left subtree (with size n_{L_i}) and T_{R_i} its right subtree (with size n_{R_i}). The size of the smaller subtree is noted by $g(i) = \min(n_{L_i}, n_{R_i})$. A search strategy $\mathcal{S}$ is said to be *self-similar* if for each decision tree, its left and right subtrees belong to $\mathcal{T}$, that is $T_{L_i} = T_{n_{L_i}}$ and $T_{R_i} = T_{n_{R_i}}$ (where trees are considered as equal if they have the same shape). And $\mathcal{S}$ is said to be *well-proportioned* if the proportion of nodes belonging to the smaller subtree approaches a limit as i tends to infinity, that is $\lim_{i\to\infty} g(i)/i$ exists. Finally we call *consistent* a strategy which is tree-growing, self-similar and well proportioned. Many usual strategies, such as linear search $(g(i)/i \to 0)$ or binary search $(g(i)/i \to 1/2)$, are consistent.

Theorem 1. *All consistent strategies are normal.*

The proof of this theorem relies on three properties of search strategies decision trees that ensure the sufficient condition for normality stated in Lemma 1.

Property 1. *For each positive integer i, the decision tree T_i has at least one external node on each unsaturated level.*

Property 2. *Let m_k be the number of decision trees with height k, the sequence $\{m_k\}_{k=1}^{\infty}$ is non-decreasing in k.*

For consistent strategies, these two properties result from self-similarity and tree-growing of the decision trees.

Property 3. *The variance of* C_n *satisfies* $s_n^2 = \Omega(n)$.

This property holds true for any decision tree, the intuition being that the tree with smallest variance is the complete binary tree (all levels saturated, except possibly the last one), which is the decision tree associated with binary search.

The proof of Theorem 1 considers two cases, $g(i)/i \to 1/2$ and $\lim g(i)/i \in [0, 1/2)$. In the first case the strategy is similar to binary search, $h_n = o(s_n)$ and the result follows from Property 3. In the second case Properties 1 and 2 are used to show that $h_n = o(s_n)$.

3. Relaxed conditions for normality

The sufficient condition for normality stated in Lemma 1 holds true for some families of tree-growing search strategies that are not consistent. For example, Fibonaccian search, where Fibonacci numbers are used to indicate the next probe, is not consistent since $\lim g(i)/i$ does not exist; but it can be shown to be normal with a proof similar to the one of Theorem 1, since $\liminf_{n\to\infty} g(n)/n$ as well as $\limsup_{n\to\infty} g(n)/n$ stand in $(0, 1/2)$.

More generally, one can exhibit different conditions on search strategies, that lead to normality by showing that the heights of the decision trees grow at a steady rate. For example

Proposition 1. *Any tree-growing search strategy* $\mathcal{S}$ *for which* $\liminf_{n\to\infty} g(n)/n$ *belongs to* $(0, 1/2)$ *is normal.*

Proposition 2. *If* $m_k = \Omega(k^{1+\epsilon})$ *for some* $\epsilon > 0$, *and* $\mathrm{Var}[X_n] = \Omega(1)$, *then the corresponding tree-growing search strategy is normal.*

Bibliography

[1] Feller (William). – *An introduction to probability theory and its applications.* – John Wiley & Sons Inc., New York, 1971, second edition, vol. II, xxiv+669p.

[2] Lent (Janice) and Mahmoud (Hosam M.). – On tree-growing search strategies. *The Annals of Applied Probability*, vol. 6, n° 4, 1996, pp. 1284–1302.