The analysis of multidimensional searching in quad-trees

Philippe Flajolet, Claude Puech, J.M. Robson, Gaston Gonnet

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.(1) 39 63 55 11

Rapports de Recherche

N° 1336

Programme 1
Programmation, Calcul Symbolique
et Intelligence Artificielle

THE ANALYSIS OF MULTIDIMENSIONAL SEARCHING IN QUAD-TREES

Philippe FLAJOLET
Gaston GONNET
Claude PUECH
J.M. ROBSON

Novembre 1990

RR.1336

The Analysis of Multidimensional Searching in Quad–Trees

Philippe Flajolet * Gaston Gonnet † Claude Puech ‡ J. M. Robson §

November 3, 1990

Abstract. *Quadtrees constitute a hierarchical data structure which permits fast access to multidimensional data. This paper presents the analysis of the expected cost of various types of searches in quadtrees—fully specified and partial match queries. The data model assumes random points with independently drawn coordinate values.*

The analysis leads to a class of "full–history" divide–and–conquer recurrences. These recurrences are solved using generating functions, either exactly for dimension $d = 2$, or asymptotically for higher dimensions. The exact solutions involve hypergeometric functions. The general asymptotic solutions relie on the classification of singularities of linear differential equations with analytic coefficients, and on singularity analysis techniques.

These methods are applicable to the asymptotic solution of a wide range of linear recurrences, as may occur in particular in the analysis of multidimensional searching problems.

L'analyse de la recherche multidimensionnelle dans les arbres Quad.

Résumé. Les arbres Quad constituent une structure de données hierarchique qui permet un accès rapide à des données multidimensionnelles. Cet article présente l'analyse du coût moyen de divers types de requêtes—recherches partiellement ou totalement spécifiées. Le modèle probabiliste est celui de points aléatoires à coordonnées réparties indépendamment.

L'analyse conduit à une classe de récurrences du type dit "diviser pour régner" avec dépendance complète de l'histoire des valeurs. Ces récurrences sont résolues par l'emploi des séries génératrices, soit exactement en dimension $d = 2$, soit asymptotiquement en dimension supérieure. Les solutions exactes mettent en jeu des fonctions hypergéométriques. Les solutions asymptotiques générales reposent sur la classification des singularités des équations différentielles linéaires à coefficients analytiques, ainsi que sur les techniques d'analyse de singularité.

Ces méthodes sont applicables à la résolution asymptotique d'une large classe de récurrences linéaires, telles celles qui se présentent dans l'analyse de la recherche multidimensionnelle.

*INRIA, Rocquencourt, F–78150 Le Chesnay, France.

†Informatik, E.T.H. Zentrum, CH–8092 Zurich, Switzerland

‡LIENS, Ecole Normale Supérieure, 45 rue d'Ulm, F–75005 Paris, France

§Department of Computer Science, Australian National University, Canberra ACT 2601, Australia.

The Analysis of Multidimensional Searching in Quad–Trees

Philippe Flajolet * Gaston Gonnet † Claude Puech ‡ J. M. Robson §

Abstract. Quadtrees constitute a hierarchical data structure which permits fast access to multidimensional data. This paper presents the analysis of the expected cost of various types of searches in quadtrees—fully specified and partial match queries. The data model assumes random points with independently drawn coordinate values.

The analysis leads to a class of "full–history" divide–and–conquer recurrences. These recurrences are solved using generating functions, either exactly for dimension $d = 2$, or asymptotically for higher dimensions. The exact solutions involve hypergeometric functions. The general asymptotic solutions relie on the classification of singularities of linear differential equations with analytic coefficients, and on singularity analysis techniques.

These methods are applicable to the asymptotic solution of a wide range of linear recurrences, as may occur in particular in the analysis of multidimensional searching problems.

*INRIA, Rocquencourt, F–78150 Le Chesnay, France.

†Informatik, E.T.H. Zentrum, CH–8092 Zurich, Switzerland

‡LIENS, Ecole Normale Supérieure, 45 rue d'Ulm, F–75005 Paris, France

§Department of Computer Science, Australian National University, Canberra ACT 2601, Australia.

Communication presented at the Second Annual ACM–SIAM Symposium on Discrete Algorithms, San Francisco, January 1991.

1 Introduction

A classical geometrical search problem consists in finding all records (points, elements) in a collection of multidimensional data (see Samet's book [20] or general references like [3, 10, 13, 16, 21]). The elements to be retrieved may be specified by several (or all) of their components. If all components are specified in the search, the problem is called a *fully specified* search. Otherwise, we call it a *partial match* query.

The *quadtree* structure is due to Finkel and Bentley [7]. It can be used to answer both fully specified and partial match search problems, and it is based on a tree data structure that extends the classical idea of a binary search tree to multidimensional data. The principle, in the case of planar problems, is simply that a point partitions the search space into four quadrants (see Fig. 1). When used recursively, this principle leads to a decomposition of the underlying search space into rectangular cells (see Fig. 2). A closely related multidimensional tree structure is the k–d tree of Bentley [2].

This paper proposes a thorough analysis of the performances of various types of searches performed on quadtrees built from "random" data. A classical framework of analysis is that of "independent" data, with components of records being independently drawn from some continuous distribu-

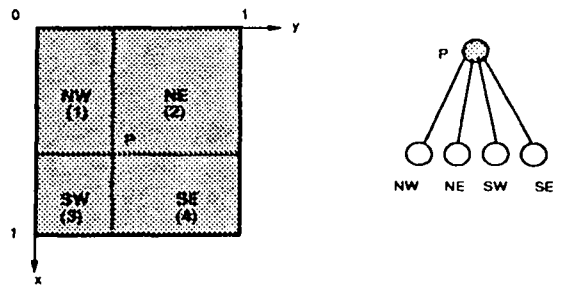

Figure 1. A point $P = (x, y)$ separates the unit square into four quadrants *NW, NE, SW, SE*, also numbered $1, 2, 3, 4$.

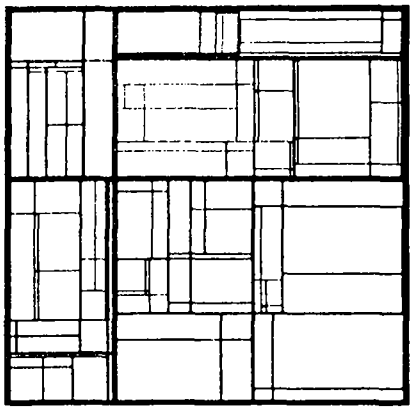

Figure 2. A quadtree decomposition of the unit square using the principle of Fig. 1 recursively, based on 50 points.

tion which we may then freely assume to be the uniform distribution over $[0, 1]$.

The quadtree is expected to provide "fast" access properties, and in particular logarithmic cost access to fully specified searches. For instance, in their original paper [7, Table 1], Finkel and Bentley observed by simulations that, for trees of size $n = 1000$ or 10000, the average cost of a search tends to be about $(0.90 \pm 0.05) \log n$. Gonnet proposes empirical formulæ implying $C_n \sim (0.989 \pm 0.004) \log n$ (for dimension $d = 2$) and $C_n \sim (0.662 \pm 0.003) \log n$ (for $d = 3$).

Our asymptotic complexity results are valid for every dimension $d \geq 2$. They are expressed in terms of the number of nodes traversed in a search, more complex measures being amenable to similar analysis techniques. A fully specified search is found to have average cost

$$C_n^{(d)} \sim \frac{2}{d} \log n. \tag{1}$$

(These results are thus in good agreement with the empirical estimates mentioned above.) If we compare the cost of a search in a common (1-dimensional) binary search tree [14] which is $\sim 2 \log n$, we thus witness a "contraction factor" of $1/d$ for the depth of d-dimensional quadtrees. This represents a sort of global conservation of the search costs (each node in a quadtree contains d fields), a phenomenon independently established by Devroye [6] using the theory of branching processes. Precise estimates with quantitative error terms appear in Theorems 1 and 4.

One of the main uses of quadtrees is for partial match queries. In that case, only s out of d coordinates, with $1 \leq s < d$, are specified in a search. First, one may consider a simplified model based on the assumption that the quadtree is a perfect tree. (See [4] and [2, p. 513] for a similar model of k-d trees.) This leads for the cost $\hat{Q}_n^{(s,d)}$ in the perfect tree model to consider the recurrence

$$\hat{Q}_n^{(s,d)} = 1 + 2^{d-s} \hat{Q}_{n/2^d}^{(s,d)}, \tag{2}$$

since a search in a tree of size n first visits the root and then continues to explore 2^{d-s} trees each of size about $n/2^d$ by the assumption of a perfect tree. The solution of Eq. (2) is

$$\hat{Q}_n^{(s,d)} = \Theta(n^{1-s/d}). \tag{3}$$

Otherwise said, a perfect quadtree resembles a perfect grid with meshes of size $n^{-1/d}$.

It turns out that the model (2) provides an undully optimistic estimate for random data. The exact form of the recurrence for the average search cost $Q_n^{(s,d)}$ is given in Section 2 below. The corrected form of (3) is then found to be

$$Q_n^{(s,d)} = \Theta(n^{1-s/d+\theta(s/d)}), \tag{4}$$

where the correction function $\theta(x)$ in the *exponent* is given in the statement of Theorem 5. For instance, when $d = 2$, a partial match query with one component specified out of two has expected cost

$$Q_n = O(n^{(\sqrt{17}-3)/2}) \approx O(n^{0.56155})$$

as opposed to $O(\sqrt{n})$ which is suggested by the approximate model. This situation resembles the

case of k-d trees which has been treated earlier by Flajolet and Puech [9], though the multiplicative constants are naturally different.

The analysis problems that we discuss here start with what may be called *stochastic divide-and-conquer* recurrences. These recurrences on average costs are direct reflections of the recursive search procedures. A typical instance is the recurrence corresponding to a fully specified search into a standard quadtree,

$$f_n = 1 + \sum_{k=0}^{n-1} \xi_{n,k} f_k, \tag{5}$$

where the $\xi_{n,k}$ are related to "splitting probabilities" (see below):

$$\xi_{n,k} = \frac{4k}{n^2}[H_n - H_k] \quad \text{with} \quad H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

The natural approach to recurrences of the form (5) is of course to introduce generating functions. We thus set

$$f(z) := \sum_{n \geq 0} f_n z^n.$$

A recurrence of the form (5), with the $\xi_{n,k}$ that involve harmonic numbers, then translates into a linear integral equation itself equivalent to a *linear differential equation* of order 2. More generally, problems in dimension d lead to differential equations of order d. The analysis of quadtrees then follows two different routes.

In dimension $d = 2$, the differential equations that we encounter have explicit solutions which, for partial match, involve *hypergeometric functions*. In this way, explicit forms—involving harmonic numbers or binomial coefficients—are available for the complexity analysis of standard quadtrees. Asymptotic forms then follow by elementary asymptotic analysis.

In dimension $d \geq 3$, we no longer find explicit forms of generating functions that would be expressible in terms of known special functions. We then follow a route inspired by the corresponding analysis of k-d trees in [9]. The principles on which the analysis is based are:

(i) The nature and location of singularities of a function determine the growth of its coefficients (see e.g., [8]).

(ii) Singularities of the solution to a linear differential equation

$$\sum_{j=0}^{d} \lambda_j(z) \frac{d^j}{dz^j} f(z) = a(z),$$

arise from singularities of the coefficients $\lambda_j(z)$ and the zeros of $\lambda_d(z)$ in a well quantified way.

The k-d trees lead to differential systems while quadtrees introduce more naturally integro-differential equations. However, in both cases, the analysis of generating functions' singularities via differential systems constitutes a fairly general methodology which may be used in order to analyze linear recurrences with coefficients that involve multiple summations and rational functions of indices.

Coming back to quadtrees, we thus establish here that, not too surprisingly, their expected performances are, as far as orders of growths are concerned, rather close to those of k-d trees. We may also mention that analyses of quadtrees under different usage, like for representing images or as an access method for databases, have been given by Yahia *et al.* [18, 25, 15] and Régnier [19].

2 Basic Probabilities and Recurrences

The average case complexity of divide-and-conquer algorithms is normally expressed by recurrences. For instance, the average number of comparisons C_n needed to sort n data items using the *Quicksort* algorithm satisfies the recurrence [14, Eq. 5.2.2-18, p. 120]

$$C_n = n + 1 + \frac{2}{n} \sum_{k=0}^{n-1} C_k, \tag{6}$$

and a closely related recurrence [14, Eq. 6.2.2-4, p. 427] provides the average search cost in a binary search tree of size n. Digital searching leads to recurrences of a different shape, see for instance [14, Eq. 6.3-17, p. 499].

The general scheme which covers the examples above as well as the quadtree costs is

$$f_n = a_n + \sum_{k=0}^{n-1} \xi_{n,k} f_k. \tag{7}$$

Problem	a_n	$\xi_{n,k}$
Quicksort	$n+1$	$\frac{2}{n}$
Binary Search	$\frac{2n}{n+1}$	$\frac{1}{n+1}$
Patricia Search	1	$\frac{1}{2^n-2}\binom{n}{k}$
Quadtree Path Length	n	$\frac{4}{n}(H_n - H_k)$
Quadtree Partial Match	1	$4\frac{n-k}{n(n+1)}$

Figure 3. Various types of stochastic divide-and-conquer recurrences. The first three recurrences appear in Knuth's book [14] (on pages 120, 427, and 479, respectively). The quadtree recurrences appear in Lemma 2 and Lemma 3.

There f_n is the unknown sequence of costs which is to be determined, a_n is a known (and usually simple) number sequence, and the $\xi_{n,k}$ are of various forms that reflect in each case the probabilities that a problem of size n decomposes into similar subproblems of size k. The form (7) is more complex than the standard divide-and-conquer recurrences of which Eq. (2) is a particular example, and we may call it a *stochastic divide-and-conquer recurrence.*

In this section, we establish the form of recurrences satisfied by the search costs in a standard quadtree of dimension $d = 2$. Let $\mathrm{U} = [0,1]^2$ denote the unit square. The probabilistic model of use[1] assumes that n elements are drawn *uniformly* and *independently* from U.

Proposition 1 *Let p_{n_1,n_2,n_3,n_4} be the probability that the four root subtrees of a quadtree built on $n = 1 + n_1 + n_2 + n_3 + n_4$ records have sizes n_1, n_2, n_3, n_4. Then:* $p_{n_1,n_2,n_3,n_4} =$

$$\frac{1}{n \cdot n!} \frac{(n_1+n_2)!\,(n_3+n_4)!\,(n_1+n_3)!\,(n_2+n_4)!}{n_1!\,n_2!\,n_3!\,n_4!}.$$

PROOF. Let $(r_1, r_2, \ldots, r_n)$ be a random element of U^n, and set $r_j = (x_j, y_j)$. The sought probability is: $p_{n_1,n_2,n_3,n_4} =$

$$\binom{n-1}{n_1, n_2, n_3, n_4} \times \int_0^1 \int_0^1 [(uv)^{n_1}((1-u)v)^{n_2} (u(1-v))^{n_3}((1-u)(1-v))^{n_4}]\,du\,dv. \tag{8}$$

There $du\,dv$ is the probability that $u \le x_1 < u+du$ and $v \le y_1 < v+dv$. The integral gives the probability that the n_1 elements, $r_2, \ldots, r_{n_1+1}$, are in the first subtree, that the next n_2 elements $r_{n_1+2}, \ldots, r_{n_1+n_2+1}$ are in the second subtree etc. Finally, the multinomial coefficient represents the number of possible "shufflings" of the $n-1$ elements $r_2, \ldots, r_n$ into four groups of cardinalities n_1, n_2, n_3, n_4.

From the classical Eulerian *Beta integral*, see [1, Chap. 6] or [23, Chap. XII], applied to Eq. (8),

$$\int_0^1 x^\alpha (1-x)^\beta\,dx = \frac{\alpha!\,\beta!}{(\alpha+\beta+1)!}, \tag{9}$$

we get the stated form of the splitting probabilities. ■

Similar arguments provide recurrences relative to path length and the cost of partial match queries.

Lemma 2 *The expected value of internal path length P_n in a random quadtree of size n satisfies the recurrence*

$$P_0 = 0; \quad P_n = n + \frac{4}{n}\sum_{k=0}^{n-1}[H_n - H_k]P_k.$$

Lemma 3 *Let Q_n be the expected value of the costs of a partial match query in a random quadtree of size n. Then Q_n satisfies the recurrence*

$$Q_0 = 0; \quad Q_n = 1 + \frac{4}{n(n+1)}\sum_{k=0}^{n-1}(n-k)Q_k.$$

3 Standard Quadtrees in Dimension $d = 2$

In this section, we carry out the analysis of search costs in standard quadtrees where the dimension is $d = 2$. Recurrences translate into integro-differential equations. For $d = 2$, the generating functions can be found explicitly. This leads both

[1] This model is of course equivalent to assuming simply independent drawings from *any* continuous distribution.

to exact and to asymptotic forms for the costs of fully specified searches and partial match queries.

In this and the next section, we use a few tools from the theory of linear differential equations for which we refer to books by Henrici [12] or Wasow [22]. A treatment of hypergeometric functions that suffices for our purposes is to be found in [1, 23].

Proposition 4 *Let* $P(z) = \sum_{n\geq 0} P_n z^n$ *and* $Q(z) = \sum_{n\geq 0} Q_n z^n$ *be the generating functions of* P_n *and* Q_n. *Then* $P(z)$ *satisfies the second order equation,* $P(0) = 0$,

$$P(z) = \frac{z}{(1-z)^2} + 4\int_0^z \frac{dt}{t(1-t)} \int_0^t P(u)\frac{du}{1-u}. \tag{10}$$

The function $Q(z)$ *satisfies the differential equation*

$$\frac{d^2}{dz^2}(zQ(z)) = \frac{2}{(1-z)^3} + \frac{4}{(1-z)^2}Q(z), \tag{11}$$

together with the initial conditions: $Q(0) = 0$, $Q'(0) = 1$.

PROOF. It follows by a direct translation from recurrences to generating functions. ■

Theorem 1 *The expected cost of a positive search in a quadtree of size* $n \geq 1$ *is*

$$C_n = \frac{P_n}{n} = (1 + \frac{1}{3n})H_n - \frac{n+1}{6n}. \tag{12}$$

The expected cost of a negative search in a quadtree of size $n \geq 1$ *is*

$$C'_n = H_n + \frac{5}{6} + \frac{1}{3(n+1)}. \tag{13}$$

PROOF. The formula for P_n was initially found by trial-and-error from exact rational forms of P_n for small n. (The occurrence of the harmonic number is not too unexpected!) Once it has been conjectured, it is a simple matter to verify that the generating function of the P_n as given by Eq. (12), namely

$$P(z) = \frac{1}{3}\frac{2z+1}{(1-z)^2}\log\frac{1}{1-z} + \frac{1}{6}\frac{z^2+4z}{(1-z)^2}, \tag{14}$$

satisfies the second order integral equation (10). ■

Corollary 5 *Asymptotically, a random search, either successful or unsuccessful, in a quadtree of size* n *has average cost* $\log n + O(1)$, *the cost being measured by the number of node traversals.*

Theorem 2 *The expected cost* Q_n *of a partial match query in a quadtree of size* $n \geq 1$ *satisfies* $1 + Q_n =$

$$\sum_{k=0}^{n} \binom{\alpha - 1 + n - k}{n-k} \binom{k-\alpha-1}{k} \binom{k-\alpha}{k} \frac{1}{k+1}, \tag{15}$$

where α *is the root located between* 1 *and* 2 *of the equation* $\alpha(\alpha+1) = 4$; *thus* $\alpha = (\sqrt{17}-1)/2 \approx 1.56155\ 28128\ 08830$.

PROOF. First, we convert the differential equation of $Q(z)$, Eq. (11), to a standard form,

$$z(1-z)^2\frac{d^2}{dz^2}Q(z) + 2(1-z)^2\frac{d}{dz}Q(z) - 4Q(z) = \frac{2}{1-z}.$$

We observe that a particular solution to this equation is $1/(1-z)$, and therefore, by considering $y(z) = Q(z) - 1/(1-z)$, we find that $y(z)$ satisfies the homogeneous equation

$$z(1-z)^2\frac{d^2}{dz^2}y(z) + 2(1-z)^2\frac{d}{dz}y(z) - 4y(z) = 0. \tag{16}$$

By general theorems, the only possible singularities of a solution to such an equation are the singularities of the coefficients, and the zeros of the leading coefficient. Thus, the only possible candidates are $z = 0$, $z = 1$, and $z = \infty$. It is known *a priori*, from the origin of the problem, that the function element $Q(z)$ is regular at 0 and has radius of convergence exactly 1 since its coefficients are polynomially bounded.

Guided[2] by the usual principles of singularity analysis, one should determine the local behaviour of $Q(z)$, or equivalently $y(z)$, around $z = 1$ in order to derive the asymptotic form of the Q_n. To that purpose, we first try to substitute an asymptotic form $y(z) \sim C/(1-z)^\alpha$ inside Eq. (16). The main terms on the left hand side of Eq. (16) are "normally" of order $(1-z)^{-\alpha}$, safe for certain exceptional values of α, where cancellation occurs through the coefficients; we expect precisely these cancellation cases to provide solutions to the differential homogeneous equation. (The left hand side

[2] See also below the paragraph on singularity analysis and in the next section the paragraph on singular differential systems.

of Eq. (16) must be identically 0.) Proceeding in this way suggests that $y(z) \sim C/(1-z)^\alpha$ with α a root of $\alpha(\alpha+1) = 4$.

To make this precise, we set

$$z\, y(z) = \frac{Y(z)}{(1-z)^\alpha}, \tag{17}$$

with α still kept as an indeterminate at the moment. The function $Y(z)$ satisfies a transformed equation, namely

$$\begin{aligned} z(z-1)^2 \frac{d^2}{dz^2} Y(z) - 2\alpha z(z-1)\frac{d}{dz} Y(z) \\ +(z\alpha^2 + z\alpha - 4)Y(z) = 0. \end{aligned} \tag{18}$$

From the preceding discussion, we now fix α to be a root of $\alpha(\alpha+1) = 4$, and we select the largest root, namely $\alpha = (\sqrt{17}-1)/2$, since it is the candidate for providing the dominant growth of $y(z)$. In so doing, a term of $(z-1)$ factors out and $Y(z)$ is found to satisfy

$$z(z-1)\frac{d^2}{dz^2} Y(z) - 2\alpha z \frac{d}{dz} Y(z) + 4Y(z) = 0. \tag{19}$$

The equation (19) clearly has three (so-called "regular") singular points at 0, 1, and ∞ and we may compare it with the standard hypergeometric equation.

The hypergeometric equation [23, p. 283] involves three parameters, a, b, c. It reads

$$\begin{aligned} z(1-z)\frac{d^2}{dz^2} F(z) + [c - (a+b+1)z]\frac{d}{dz} F(z) \\ -abF(z) = 0. \end{aligned} \tag{20}$$

A formal solution of it defines the classical *hypergeometric function*, $F[a, b; c; z] =$

$$1 + \frac{a \cdot b}{c}\,\frac{z}{1!} + \frac{a(a+1)\cdot b(b+1)}{c(c+1)}\,\frac{z^2}{2!} + \cdots. \tag{21}$$

Clearly special cases arise when parameters assume special values, e.g., if c is a negative integer.

It is now an easy task to match the hypergeometric equation (20) with the equation (19) satisfied by $Y(z)$. We find the correspondence

$$a = -\alpha, \quad b = -(\alpha+1), \quad c = 0. \tag{22}$$

The fact that $c = 0$ is an indication that we are in one of the special cases of the hypergeometric equation. It is known (see Article 15.5.20, page 564 of [1]) that one of the solutions is then $F(a+1, b+1, 2, z)$, and another independent solution has a logarithmic singularity at 0. (The series solution can also be verified directly by the method of indeterminate coefficients.)

Since $y(z)$ and $Q(z)$ are by construction regular at 0, logarithmic solutions should be discarded. The coefficient of the solution that is analytic at the origin is easily found to be equal to 1, because of initial conditions. Thus, from the correspondence of Eq. (22), unwinding our earlier changes of variables, we find the main equation

$$Q(z) = \frac{F[-\alpha, 1-\alpha; 2; z]}{(1-z)^\alpha} - \frac{1}{1-z}, \tag{23}$$

with $\alpha = (\sqrt{17}-1)/2$, the hypergeometric function F being defined by (21).

In the particular case when the parameter $c = 2$, we find

$$F[a, b; 2; z] = \sum_{n=0}^{\infty} \binom{a+n-1}{n} \binom{b+n-1}{n} \frac{z^n}{n+1}.$$

By the binomial expansion, we also have

$$\frac{1}{(1-z)^\alpha} = \sum_{n=0}^{\infty} \binom{\alpha+n-1}{n} z^n.$$

These two expansions permit us to determine an explicit convolution form of the coefficients of $Q(z)$, as obtained in (23). The statement of the theorem follows. ■

From the generating function form (23) of $Q(z)$, detailed asymptotic information on the coefficients Q_n is available. By the general principles of *singularity analysis* techniques [8] that we review now, the asymptotic form of Q_n is determined by the asymptotic properties of $Q(z)$ at its singularity $z = 1$.

Singularity analysis. That method is based on two principles. First, if we examine coefficients of standard functions that are singular at $z = 1$, we observe that functions that get larger around $z = 1$ have larger coefficients. Let $[z^n]f(z)$ denote the coefficient of z^n in $f(z)$. Approximating the binomial coefficients, we find

$$[z^n]\,\frac{1}{(1-z)^\alpha} = \frac{n^{\alpha-1}}{\Gamma(\alpha)} + O(n^{\alpha-2}). \tag{24}$$

Next, it can be proved under a variety of conditions that the type of estimate (24) also holds for

functions only known asymptotically at $z = 1$,

$$[z^n]\, O(\frac{1}{(1-z)^\beta}) = O(n^{\beta-1}). \qquad (25)$$

One set of conditions ensuring the validity of the "transfer" of (25) is that the expansion of the function holds in an extended domain of the complex plane.

The combination of (24) and (25) shows that once a singular expansion of a function has been obtained, the asymptotic form of its Taylor coefficients is known. Thus, under the analytic continuation conditions of [8], we have the implication:

$$f(z) = \frac{C}{(1-z)^\alpha} + O(\frac{1}{(1-z)^\beta})$$
$$\Longrightarrow\ [z^n]\, f(z) = \frac{C}{\Gamma(\alpha)} n^{\alpha-1} + O(n^{\alpha-2} + n^{\beta-1}).$$

Theorem 3 *The expected cost of a partial match query in a quadtree of size $n \geq 1$ satisfies asymptotically*

$$Q_n \sim \gamma n^{\alpha-1} \quad \textit{where} \quad \gamma = \frac{1}{2}\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^3}, \qquad (26)$$

with $\alpha = (\sqrt{17} - 1)/2$. Numerically $\gamma \approx$ 1.59509 90958 29715.

PROOF. First, by a classical identity of Gauß, we have

$$F[a, b; c; 1] = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}, \qquad (27)$$

whenever $\Re(c-a-b) > 0$, and $c \neq 0, -1, -2, \ldots$ (see [23, 14.11] or Article 15.1.20 of [1]). Thus, we find from Eq. (23) that

$$Q(z) \sim \frac{\gamma^*}{(1-z)^\alpha} \qquad (z \to 1), \qquad (28)$$

with

$$\gamma^* = F[-\alpha, 1-\alpha; 2; 1] = \frac{\Gamma(2\alpha+1)}{\Gamma(2+\alpha)\Gamma(1+\alpha)}.$$

That asymptotic expansion is easily found to hold true in an extended domain of the complex plane since the hypergeometric function only has algebraic or logarithmic branch points. Thus, by singularity analysis [8], we are able to "transfer" the asymptotic relation on $Q(z)$ into a corresponding asymptotic form of Q_n, namely

$$Q_n \sim \gamma^* \frac{n^{\alpha-1}}{\Gamma(\alpha)}.$$

The statement of the theorem thus follows with $\gamma = \gamma^*/\Gamma(\alpha)$. ■

A refinement of this argument leads to a full asymptotic expansion for the Q_n.

Corollary 6 *Define the asymptotic series in n,*

$$\phi(\theta, n) \sim 1 + \sum_{k=1}^{\infty} \frac{(\theta-1)^3 \cdots (\theta-k)^3}{(2\theta)\cdots(2\theta-k+1)} \frac{\theta}{\theta-k} \cdot \frac{(-1)^k}{k!} \frac{1}{(n+\theta-1)\cdots(n+\theta-k)}.$$

Then

$$1 + Q_n \sim \frac{1}{2}\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}\binom{n+\alpha-1}{n}\phi(\alpha, n) + \frac{1}{2}\frac{\Gamma(2\bar{\alpha})}{\Gamma(\bar{\alpha})^2}\binom{n+\bar{\alpha}-1}{n}\phi(\bar{\alpha}, n), \qquad (29)$$

with $\alpha = (-1+\sqrt{17})/2$, and $\bar{\alpha}$ the conjugate of α, $\bar{\alpha} = (-1-\sqrt{17})/2$.

The form (29) provides an asymptotic expansion (that is divergent!) of Q_n. The asymptotic scale involves inverses of descending "factorials" of $n+\alpha$ and $n+\bar{\alpha}$.

We have thus found a new expansion of Q_n as a sum of two purely divergent formal ${}_3F_2$–hypergeometric forms. The quality of the approximation that we obtain by retaining the first four terms of the expansion (29)—these terms all come from $\phi(\alpha, n)$—is already quite exceptional; for $n = 1, 10, 100, 1000$, the absolute error is of order respectively $10^{-2}, 10^{-6}, 10^{-9}, 10^{-12}$. The error is tiny, even for $n = 1$, and even though the series is divergent!

4 Higher Dimensions

In this section we examine the cost of various searches in quadtrees for data taken in higher dimensional spaces. The recurrences involve more complicated splitting probabilities and the generating function equations have integral forms that reduce to linear differential equations of order d, when the dimension is equal to d. The results are less explicit than in the case $d = 2$, but orders of growth can still be precisely quantified although,

in the case of partial match, the multiplicative constants do not appear to have closed forms (to the best of our knowledge!).

We use in an essential manner singularity analysis techniques. We are thus led to analyzing generating functions locally around their dominant singularity at $z = 1$.

> The case of a fully specified search illustrates a situation in which the dominant asymptotic behaviour at $z = 1$ comes from the inhomogeneous term in the differential equation.
>
> The case of a partial match query corresponds to a situation where the dominant asymptotic contribution comes from solutions to the associated homogeneous equation.

In both cases, we use a modest amount of the theory of singular points of linear differential equations as may be found in books by Henrici [12] or Wasow [22].

Singular differential systems. By a classical theorem, the singularities of a homogeneous *linear* differential equation or system can only arise from singularities of the coefficients. For systems, a particularly important case occurs when the coefficient matrix is meromorphic and the singularity under consideration is only a *simple* pole. The singularity is then called *regular*. If the singularity is normalized to occur at $z = 1$, a fundamental result implies that there exist solutions of the form

$$\frac{1}{(1-z)^{\alpha}} \cdot \sum_{k=0}^{\infty} c_k (1-z)^k.$$

By substituting inside the original equations, we need to obtain complete cancellation. It is then seen that only a finite number of possibilities exist for α; these are solutions of a polynomial equation which is known as the *indicial equation*. The process could be called a method of "indeterminate exponents"; complete expansions then follow by the usual technique of indeterminate coefficients.

Lemma 7 *Let P_n denote the expected internal path length in a d-dimensional quadtree of size n. The P_n satisfy the recurrence*

$$P_n = n + 2^d \sum_{k=0}^{n-1} \pi_{n,k} P_k \quad \textit{with}$$
$$\pi_{n,k} = \frac{1}{n} \sum_{\mathcal{L}} \frac{1}{(\ell_1+1)(\ell_2+1)\cdots(\ell_{d-1}+1)}, \tag{30}$$

where the summation is over all sequences $(\ell_1, \ell_2, \ldots, \ell_d)$ with the condition $\mathcal{L}$ being $n > \ell_1 \geq \ell_2 \geq \cdots \geq \ell_{d-1} \geq \ell_d = k$

The generating function $P(z) = \sum_{n\geq 0} P_n z^n$ satisfies the integral equation

$$P(z) = \frac{z}{(1-z)^2} + 2^d \mathbf{J}^{d-1} \mathbf{I} f(z), \tag{31}$$

where the operators $\mathbf{I}, \mathbf{J}$ are defined by

$$\mathbf{I}f(z) = \int_0^z f(t) \frac{dt}{1-t},$$
$$\mathbf{J}f(z) = \int_0^z f(t) \frac{dt}{t(1-t)}.$$

Theorem 4 *The expected cost $C_n^{(d)}$ of a fully specified search in a d-dimensional quadtree of size n satisfies:* $C_n^{(d)} =$

$$\frac{2}{d} \log n + \lambda_d + O\left(\frac{\log n}{n} + n^{2\cos(2\pi/d)-2} \log n\right), \tag{32}$$

for some real constant λ_d.

PROOF. The main idea of the proof is that the equation (31) behaves as a perturbation of a simpler equation that can be solved explicitly. This fact relies on the observation that the two functionals $\mathbf{I}f(z)$ and $\mathbf{J}f(z)$ act as "singularity transformers" (around the singularity $z = 1$) in the same way, as far as main orders of growth are concerned. The proof proceeds in three steps.

A. Consider the simplified homogeneous equation in which $\mathbf{J}$ is replaced by $\mathbf{I}$,

$$y(z) - 2^d \mathbf{I}^d y(z) = 0. \tag{33}$$

This is an Euler equation that has exact solutions of the form $(j = 0, \ldots, d-1)$

$$y_j(z) = (1-z)^{-2\omega^j} \qquad \text{with} \qquad \omega = e^{2i\pi/d}. \tag{34}$$

B. The inhomogeneous equation associated with (33) and (31) is

$$g(z) - 2^d \mathbf{I}^d g(z) = \frac{z}{(1-z)^2}. \tag{35}$$

Put differently, the function $g(z)$ is a component of a vectorial system,

$$\frac{d}{dz}\mathbf{g} = \frac{2}{1-z}\mathbf{A}\mathbf{g} + \mathbf{w}, \tag{36}$$

in which the matrix $\mathbf{A}$ involved in the singular part is a circular *permutation matrix*. The inhomogeneous system is then solved by the matrix form of the variation–of–constant method [12, p. 99], since all solutions to the homogeneous equation are known.

In this way, we find that the solution to (35) satisfies

$$g(z) = \frac{2}{d}\frac{1}{(1-z)^2}\log\frac{1}{1-z} + \frac{C}{(1-z)^2} + o\left(\frac{1}{(1-z)^2}\right). \tag{37}$$

The logarithm occurs because of "resonances" between some homogeneous solutions and the inhomogeneous term. Notice carefully that we are also able to determine ultimately the leading coefficient $2/d$ because the treatment of (36) can be made explicit.

C. Let $g(z)$ be the solution to the approximate inhomogeneous equation (35). We return to the equation

$$f(z) - 2^d \mathbf{J}^{d-1}\mathbf{I} f(z) = \frac{z}{1-z}, \tag{38}$$

satisfied by $P(z)$. The induced equation for $h(z) = f(z) - g(z)$ is such that its homogeneous part resembles that of the original equation while we attain a reduction in order of growth for the inhomogeneous term. In this way, we are able to prove that $h(z) = o((1-z)^{-2})$. Thus the dominant asymptotic growth is dictated by $g(z)$, and the singular expansion of $P(z)$ has been completed.

From there, the behaviour of P_n, hence that of $C_n^{(d)} = P_n/n$, follows by singularity analysis. ∎

Lemma 8 *Let Q_n represent the expected number of node traversals in a partial match query of a random d-dimensional quadtree containing n points. Then $Q_0 = 0$ and, for $n \geq 1$, the Q_n satisfy the recurrence,*

$$Q_n = 1 + 2^d \sum_{k=0}^{\infty} \pi^*_{n,k} Q_k \tag{39}$$

with $$\pi^*_{n,k} = \frac{1}{n(n+1)}\left[\frac{1}{(\ell_1+2)\cdots(\ell_{s-1}+2)}\right]\cdot\left[\frac{1}{(\ell_{s+1}+1)\cdots(\ell_{d-1}+1)}\right],$$

where the summation takes place over all sequences $(\ell_1, \ell_2, \ldots, \ell_d)$ satisfying the condition $\mathcal{L}$ above.

The generating function $Q(z) = \sum_{n\geq 0} Q_n z^n$ satisfies the integral equation,

$$z\frac{d^2}{dz^2}(zQ(z)) = \frac{2z}{(1-z)^3} + 2^d \frac{1}{1-z}\mathbf{J}^{s-1}\frac{z}{1-z}\mathbf{I}^{d-s-1}Q(z). \tag{40}$$

That lemma provides the basic recurrences that hold true for the higher dimensional partial match search. Its proof is based on an extension of the methods employed for the computation of geometric probabilities in Section 2.

Theorem 5 *The expected cost $Q_n^{(s,d)}$ of a partial match query in a d-dimensional quadtree of size n, with s coordinates specified, satisfies asymptotically*

$$Q_n^{(s,d)} \sim \gamma_{s/d} n^{\alpha-1} \tag{41}$$

for some constant $\gamma_{s,d} \neq 0$ and α the root between 1 and 2 of the equation

$$\alpha^{d-s}(\alpha+1)^s = 2^d.$$

In other words, we have

$$Q_n^{(s,d)} \sim \gamma_{s/d} n^{1-s/d+\theta(s/d)},$$

where the function $\theta(u)$ is defined as the solution $0 < \theta < 0.07$ of the equation

$$(\theta+3-x)^x(\theta+2-x)^{1-x} - 2 = 0. \tag{42}$$

PROOF. The linear integral equation (40) is transformed into a differential equation of order d by repeatedly taking derivatives.

The homogeneous equation has d linearly independent solutions, each of the form

$$\frac{1}{(1-z)^{\alpha_j}} \sum_{k=0}^{\infty} c_k^{(j)}(1-z)^k, \tag{43}$$

for $j = 1, \ldots, d$. The α_j are determined as the roots of the indicial equation

$$\alpha^{d-s}(1+\alpha)^s = 2^d. \tag{44}$$

(It can be shown that this polynomial has all its roots that are distinct.)

Amongst the functions (43), it is the one with $\Re(\alpha_j)$ maximal that gives the dominant contribution around $z = 1$, amongst all solutions to the homogeneous equation. This corresponds to the unique solution α of the indicial equation (44) that belongs to the real interval $(1, 2)$.

It can be shown that, contrary to the case of a full search, the inhomogeneous terms introduce contributions that are asymptotically negligible at $z = 1$. In summary, we have found that

$$Q(z) \sim \frac{C}{(1-z)^{\alpha}}, \quad \alpha^{d-s}(1+\alpha)^s = 2^d, \ \alpha \in (1,2). \tag{45}$$

That asymptotic form can then be transferred to Q_n by the usual methods of singularity analysis. ■

Conclusions. Multidimensional search problems may lead to intricate divide-and-conquer recurrences. A method of great generality consists in studying these recurrences via generating functions.

Recurrences of the full history type with coefficients that are rational functions of n and k lead to linear integro-differential systems. These systems, once transformed into linear differential equations (with analytic coefficients) can be analyzed *locally* in the neighbourhood of their singularities, using the classical results from the theory of linear differential equations. The singular expansions so obtained can then be "transferred" back to the original number sequence, by means of the method of singularity analysis.

We thus have available a method of considerable generality[3] that can be used to study a large number of linear recurrences (with variable coefficients!), both in homogeneous and inhomogeneous cases.

The techniques of this paper have been applied recently by Flajolet and Hoshi for determining the storage occupation of paged quadtrees. They can also be used [17, p. 174] in order to analyze the average cost of finding the records with smallest x-coordinate: for instance, for dimension 2, the cost is of the form $O(n^{\sqrt{2}-1})$, see Guibas' problem in the *Journal of Algorithms* [11]. (A somewhat related problem of geometric probabilities is also discussed in [5].) In general dimension d, the expected cost of finding the point with minimum x-coordinate in quadtrees is found to be of the form

$$O(n^{-1+2^{1/d}}).$$

Quite clearly, any suitably "additive" parameter of quadtrees in dimension ≥ 2 can be analyzed by the methods described here.

Acknowledgement. Work of the first author was supported in part by the Basic Research Action of the E.C. under contract No. 3075 (Project ALCOM).

[3] That method constitutes an alternative to the direct treatment of recurrences by the theory of difference equations that was developed by G. D. Birkhoff, and is discussed extensively by Wimp and Zeilberger in [24].

References

[1] Abramowitz, M., and Stegun, I. A. *Handbook of Mathematical Functions.* Dover Publications, 1973.

[2] Bentley, J. L. Multidimensional binary search trees used for associative searching. *Communications of the ACM 18*, 9 (September 1975), 509–517.

[3] Bentley, J. L., and Friedman, J. H. Data structures for range searching. *ACM Computing Surveys 11*, 4 (1979), 397–409.

[4] Bentley, J. L., and Stanat, D. F. Analysis of range searching in quad trees. *Information Processing Letters 3*, 6 (July 1975), 170–173.

[5] Buchta, C. On the average number of maxima in a set of vectors. *Information Processing Letters 33* (November 1989), 63–65.

[6] Devroye, L. Branching processes in the analysis of the heights of trees. *Acta Informatica 24* (1987), 277–298.

[7] Finkel, R. A., and Bentley, J. L. Quad trees, a data structure for retrieval on composite keys. *Acta Informatica 4* (1974), 1–9.

[8] Flajolet, P., and Odlyzko, A. M. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics 3*, 2 (1990), 216–240.

[9] Flajolet, P., and Puech, C. Partial match retrieval of multidimensional data. *Journal of the ACM 33*, 2 (1986), 371–407.

[10] GONNET, G. H. *Handbook of Algorithms and Data Structures.* Addison-Wesley, 1984.

[11] GUIBAS ED., L. Problems. *Journal of Algorithms 3*, 4 (1982), 362-380. (Problem 80-6), from the Stanford 1979 Algorithms Qualifying Examination. Solution by Eric S. Rosenthal, pp. 368-371.

[12] HENRICI, P. *Applied and Computational Complex Analysis*, vol. 2. John Wiley, New York, 1977.

[13] IYENGAR, S. S., RAO, N. S. V., KASHYAP, R. L., AND VAISHNAVI, V. K. Multidimensional data structures: Review and outlook. *Advances in Computers 27* (1988), 69-119.

[14] KNUTH, D. E. *The Art of Computer Programming*, vol. 3: Sorting and Searching. Addison-Wesley, 1973.

[15] MATHIEU, C., PUECH, C., AND YAHIA, H. Average efficiency of data structures for binary image processing. *Information Processing Letters 26* (October 1987), 89-93.

[16] PREPARATA, F. P., AND SHAMOS, M. I. *Computational Geometry, An Introduction.* Springer Verlag, 1985.

[17] PUECH, C. *Méthodes d'analyse de structures de données dynamiques.* Doctorat ès sciences, Université de Paris Sud, Orsay, 1984.

[18] PUECH, C., AND YAHIA, H. Quadtrees, octrees, hyperoctrees: a unfied approach to tree data structures used in graphics, geometric modeling and image processing. In *First ACM Symposium on Computational Geometry* (Baltimore, 1985), pp. 272-280.

[19] RÉGNIER, M. Analysis of grid file algorithms. *BIT 25* (1985), 335-357.

[20] SAMET, H. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley, 1990.

[21] SEDGEWICK, R. *Algorithms*, second ed. Addison-Wesley, Reading, Mass., 1988.

[22] WASOW, W. *Asymptotic Expansions for Ordinary Differential Equations.* Dover Publications, 1987.

[23] WHITTAKER, E. T., AND WATSON, G. N. *A Course of Modern Analysis*, fourth ed. Cambridge University Press, 1927. Reprinted 1973.

[24] WIMP, J., AND ZEILBERGER, D. Resurrecting the asymptotics of linear recurrences. *Journal of Mathematical Analysis and Applications 111* (1985), 162-176.

[25] YAHIA, H. ***Analyse des structures de données arborescentes représentant des images.*** Doctorat de troisième cycle, Université de Paris Sud, Orsay, December 1986.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249-6399