Partial match retrieval of multidimensional data

Philippe Flajolet, Claude Puech

INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. : (3) 954 90 20

Rapports de Recherche

N° 233

PARTIAL MATCH RETRIEVAL OF MULTIDIMENSIONAL DATA

Philippe FLAJOLET
Claude PUECH

Août 1983

PARTIAL MATCH RETRIEVAL OF MULTIDIMENSIONAL DATA

Philippe FLAJOLET and Claude PUECH***

Abstract : We present a precise analysis of partial match retrieval of multi-dimensional data. The structures considered here are multidimensional search (trees (k-d-trees) dans digital search trees (k-d-tries) as well as structures designed for efficient retrieval of information stored on external devices. The methods used include a detailed study of a differential system around a regular singular point in conjonction with suitable contour integration techniques (for the analysis of k-d-trees) and properties of the Mellin integral transform (for k-d-tries) and extendible cell algorithms).

Résumé : Nous présentons une analyse précise de la recherche partiellement spécifiée de données multidimensionnelles. Les structures étudiées sont les arbres binaires de recherche et les arbres lexicographiques multidimensionnels ("k-d-trees", "k-d-tries") ainsi que certaines structures ("grid-files", "extendible cells") conçues pour une recherche efficace en mémoire externe. Les analyses reposent, en particulier, sur une étude précise du comportement des solutions d'un système différentiel autour d'un point singulier régulier en liaison avec des techniques d'intégration de contour, et sur l'utilisation de propriétés de la transformation de Mellin.

* *I.N.R.I.A.*
78150 ROCQUENCOURT (France)

** *Laboratoire de Recherches en Informatique*
ERA 452 "Al Khowarizmi" du CNRS
Université de Paris-Sud, Bât. 490
91405 ORSAY (France)

and

Ecole Normale Supérieure
1 rue Maurice Arnoux
92120 MONTROUGE (France)

INTRODUCTION

Methods for retrieval of multidimensional data are of prime importance to the design of data base systems and to specific applications including the management of geographical data or graphics algorithms. The ancestry of most currently developped algorithms is to be found in early works by Rivest [Ri76], where hashing and digital techniques are explored, and by Bentley, Finkel [Be75], [FB74] who proposed quad-tree and k-d-trees which are comparison-based structures. A description of early algorithms appears in section 6.5 of Knuth's book [Kn73]. Recent developments in the context of large external files combine some of these techniques with ideas derived from dynamic hashing schemes for single-attribute records (Virtual Hashing [Li78], Dynamic Hashing [La78], Extendible Hashing [FNPS79]); a few such examples are the grid-file of Nievergelt et al. [NHS81], the extendible cell method [Ta82] and the multidimensional extendible hashing algorithm of [LR82].

This paper describes evaluation methods for the major multidimensional search algorithms. We concentrate here on the problem known as partial match retrieval where all records in a file having specified values for some of their attributes are to be found. Contrary to the case of single attribute search, no general algorithm is known for locating a record in a file of size or using O(logn) time and linear storage (It is indeed conjectured that no such algorithm exists, see [Ri76]). We prove here that the average search cost in a file of size n, containing k-dimensional records, when s attributes are specified ($0<s<k$) is:

(a) for k-d-trees: $O(n^{1-s/k+\theta(s/k)})$ field comparisons where $\theta(u)$ is a strictly positive function of u for $0<u<1$, with maximum value 0.07. This result is of interest since it disproves an (often quoted) old claim of Bentley that k-d-trees perform in expected time $O(n^{1-s/k})$.

(b) for k-d-tries: $O(n^{1-s/k})$ bit comparisons where the implied constant in the $O(\ )$ is precisely characterized and turns out to be quite small. This result is a useful complement to some of Rivest's analyses made under a different statistical model which suggested a higher order of $O(n^{\log_2(2-s/k)})$ for k-d-tries.

(c) for grid-file algorithms: $O(n^{1-s/k})$ page accesses; there again the implied constants can be precisely determined.

A comparison of these results shows that, for multidimensional search trees, digital methods asymptotically outperform comparison-based techniques. As an example, partial match retrieval of 2-dimensional records with one attribute specified has average cost:

$$O(n^{\frac{\sqrt{17}-3}{2}}) = O(n^{0.56}) \text{ for 2-d-trees}$$

and $O(n^{1/2})$ for 2-d-tries.

Performances of the type $O(n^{1-s/k})$ have been conjectured to be optimal by Rivest [Ri76].

We feel that the interest of the paper also lies in the proof techniques employed (especially in case (a), where previous analyses appear to be invalid):

(a)- For k-d trees, we start by setting up a system of integral equations for adequately chosen generating functions of costs. The system transforms into a linear differential system (with variable coefficients) of order 2k-s, which does not seem to admit closed form solutions. Indeed, the shape of our final results strongly suggests that no such form exists and that no elementary combinatorial approach is likely to be workable. We then proceed to study the way the system becomes singular, and with the help of classical results from the theory of "regular singular points" of differential systems, we obtain the asymptotic behaviour of cost generating functions around their common singularity. We then use the Cauchy integral formula for Taylor coefficients of power series in conjunction with suitable contours of integration (in a manner similar to [FO82]) to conclude the analysis of k-d-trees.

(b-c) There, we set up in each case a system of difference equations for generating functions of costs that can be solved explicity. This leads to exact expressions for the average case behaviour of algorithms considered. We then appeal to Mellin transform techniques (see [Kn73])† to derive the results stated in (b) and (c) relative to k-d-tries and grid-file algorithms.

It should be stressed that the methods used here are of a rather wide applicability: those of type (a) could serve to derive direct asymptotic evaluations for a number of comparison based algorithms; methods of type (b-c) may be used to analyze in detail a number of data structures and algorithms closely related to tries, like the double chained trees [CS 77], multiattribute trees [KSY77] and the like. These analysis will be given in a companion paper (see [FP83]. For a preliminary report).

† See in particular the Section on radix exchange sort [Kn73,p 131] where Knuth uses Mellin transform techniques under the name of "Gamma function method".

1- GENERAL SETTING

We consider the problem of retrieving multiattribute records that belong to some k-dimensional domain

$$D = D_1 \times D_2 \times \ldots \times D_k$$

A file F is any finite subset of D and the size of F, usually denoted by n in the sequel is the number of elements in F. Our interest is in data structures for performing partial match retrieval: given F and a query $q=(q_1,q_2,\ldots q_k)$:

$$q \in (D_1 \cup \{*\}) \times (D_2 \cup \{*\}) \ldots \times (D_k \cup \{*\}),$$

one is asked to find all records in F satisfying query q, i.e. to determine the subset q(F) of F of records $r=(r_1,r_2,\ldots r_k)$ in F satisfying for all j: $1 \leq j \leq k$

$$r_j = q_j \quad \text{if} \quad q_j \neq *$$

Thus a query q= (TOTO, * , 39, 35000, *) asks for all (5-dimensional) records whose first attribute is TOTO, third attribute 39 and fourth attribute 35000, attributes 2 and 5 being left unspecified. The specification pattern of a query q is a word u of length k over the alphabet {S,*} where $u_j = S$ if q_j is specified and $u_j = *$ if q_j is left unspecified. In the above example, the specification pattern is thus S*SS*.

In the sequel, for the sake of unity, we assume that each of the attribute domain is assimulated to the real interval [0;1]; this is practically justified when the binary encodings of attributes are sufficiently long strings. Our analyses are relative to the uniform probabilistic model where we assume that attributes in either files or queries are uniformly and independently distributed over the interval. As is well known, in the case of comparison based algorithms, this model is equivalent to the more general model where attributes are only assumed to be independently drawn from any continuous distribution over any interval, so that there the uniform model is general enough. In the case of digital techniques, the uniform model constitutes an excellent approximation to real situations when superposed hashing is used, and otherwise an optimistic model of varying accuracy depending upon the particular structure of the data manipulated; however, our analyses can be easily generalized to cover biased probabilities of occurrences of bits or characters in records, and the orders of magnitude of expected case complexities appear to be only very slightly affected by this change in the model. Thus our general conclusions remain valid for a wide range of situations.

The general pattern of our analyses is as follows: we let $c_{u,n}$ with n an integer and $u = u_1, u_2 \ldots u_k$ a specification pattern denote the expected cost of a query with specification pattern u in a file of size n. We then introduce some generating function $c_u(z)$ of the sequence $\{c_{u,n}\}_{n\geq 0}$. We find in each case (a), (b), (c) that there are two operators Φ_* and Φ_S such that

$$\Sigma \quad \begin{cases} c_u(z) = \Phi_{u_1}(c_{u'}(z)) \\ c_{u'}(z) = \Phi_{u_2}(c_{u''}(z)) \\ \cdot \\ \cdot \\ \cdot \\ c_{u^{(k-1)}}(z) = \Phi_{u_k}(c_u(z)) \end{cases}$$

where u',u",u"',... designate the patterns obtained by circularly shifting the letters of u to the left by 1,2,3,... positions. The structure of system Σ reflects the cyclical changes of the partitioning attributes in the multidimensional trees.

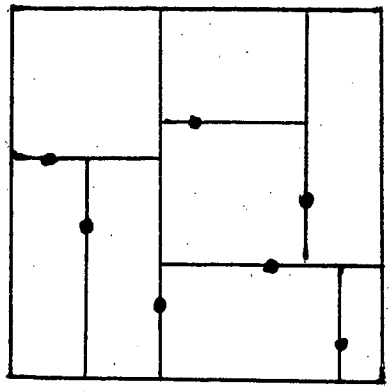
For k-d-trees, Φ_* and Φ_S turn out to be integral operators; for the other cases, they are difference operators.

2- MULTIDIMENSIONAL BINARY SEARCH TREES

Multidimensional binary search trees (or k-d-trees) are constructed by repeated insertions from the file to be represented. At the root of the tree, we use the first field of the record stored there as a discriminator; we choose to go right or left by comparing the first field of the record to be inserted with the first field of the root (going to the left if it is smaller, going to the right otherwise). At the second level of the tree, the second attribute serves to discriminate records and so on, attributes 1,2,3,...,k being used cyclically as discriminators. From the definition follows that 1-d trees coïncide with the usual binary search trees.

A partial match query proceeds along the tree, branching to one side if the corresponding field is specified by the query or proceeding along both subtrees if the field is unspecified.

From the definition also follows that a k-d-tree can be viewed as a recursive partitioning of the underlying space according to alternate dimensions. Figure 1 represents a tree constructed from a file of 7 elements together with the associated partitioning of the plane.



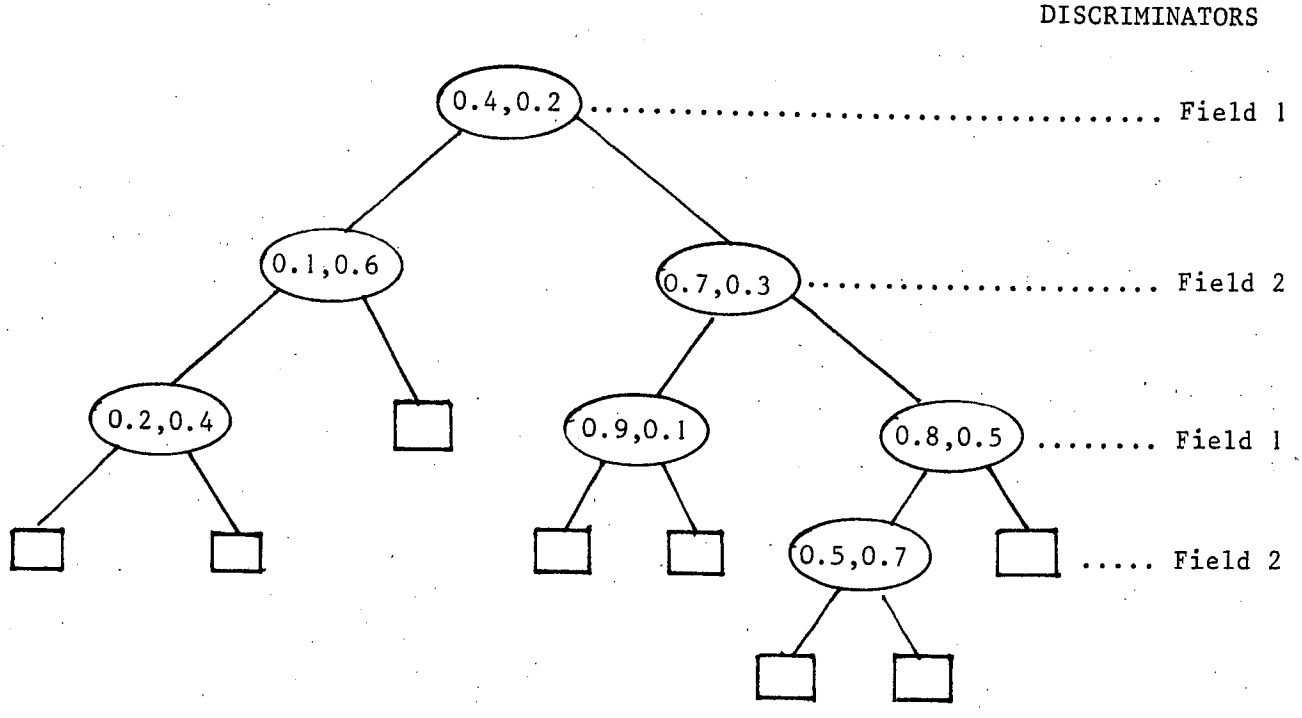


Figure 1: The 2-d-tree associated to the file F= {(0.4,0.2); (0.1,0.6); (0.7,0.3); (0.2,0.4); (0.8,0.5); (0.9,0.1); (0.5,0.7)}, elements arriving in the order they are listed, and a representation of the corresponding partitioning of [0;1] x [0;1].

The main theorem that we prove for k-d-trees is as follows:

Theorem 1: The average cost of a partial match query of specification pattern u in a k-d-tree constructed by random insertions from a file of size n satisfies:

$$c_{u,n} = \gamma_u \, n^{1-s/k+\theta(s/k)} \, [1+o(1)]$$

where γ_u is a strictly positive real constant, and the function $\theta(x)$ is defined as the unique positive real root in the interval $[0;1]$ of the equation:

$$(\theta(x)+3-x)^x \quad (\theta(x)+2-x)^{1-x} - 2 = 0,$$

so that, for $0<x<1$:

$$0<\theta(x) < 0.07$$

This theorem is proved through a chain of lemmas. Lemma 1 below expresses the recurrences satisfied by the quantities $c_{u,n}$, $c_{u',n}$, $c_{u'',n}$ with u,u',u"... being the successive left circular shifts of u. The natural expressions of these recurrences is in terms of corresponding generating fonctions.

Lemma 1: For each specification pattern u, define the generating functions:

$$c_u(z) = \sum_{n\geq 0} c_{u,n} z^n$$

$$d_u(z) = \sum_{n\geq 0} c_{u,n}(n+1) z^n.$$

One has:

(i) if u= *v (i.e. the first attribute is unspecified):

$$c_u(z) = \frac{1}{1-z} - 1+2\int_0^z c_{u'}(t) \frac{dt}{1-t}$$

(ii) if u= Sv (i.e. the first attribute is specified):

$$d_u(z) = \frac{1}{(1-z)^2} - 1+2\int_0^z d_{u'}(t) \frac{dt}{1-t} .$$

Proof: (i) The average search cost in a fixed tree $t=t_1^{\frown}t_2$ satisfies:

$$c_u[t] = 1+c_{u'}[t_1] + c_{u'}[t_2], \qquad (1)$$

since, the first attribute being unspecified, one needs to visit the root of the tree and then recursively continue the search in both t_1 and t_2 with speci-

fication pattern u'. Taking expected values of (1), and noticing that the probability that t_1 contains p nodes, for any p: $0 \le p < n$, is uniformly $\frac{1}{n}$ (thus being independent of p), we find, for $n \ge 1$:

$$c_{u,n} = 1 + \frac{1}{n} \sum_{p=0}^{n-1} [c_{u',p} + c_{u',n-1-p}],$$

and by symmetry:

$$c_{u,n} = 1 + \frac{2}{n} \sum_{p=0}^{n-1} c_{u',p} \qquad (2)$$

Taking corresponding generating functions, and using (2), establishes part (i) of the claim of the lemma.

(ii) The average search cost in a fixed tree $t = \overset{}{\underset{t_1}{\wedge}}\!\!\overset{}{\underset{t_2}{}}$ when the first attribute is specified satisfies:

$$c_u[t] = 1 + \frac{p+1}{n+1} c_{u'}[t_1] + \frac{n-p}{n+1} c_{u'}[t_2], \text{ where } p = |t_1|. \qquad (3)$$

This correspond to the fact that a search with first attribute specified proceeds along t_1 with probability $\frac{p+1}{n+1}$ and along t_2 with the complementary probability. Thus multiplying (3) by (n+1), and taking average values over all possible trees t, we get in a similar manner for $n \ge 1$:

$$(n+1)\, c_{u,n} = (n+1) + \frac{1}{n} \sum_{p=0}^{n-1} [(p+1)\, c_{u',p} + (n-p)\, c_{u',n-1-p}]$$

$$(n+1)\, c_{u,n} = (n+1) + \frac{2}{n} \sum_{p=0}^{n-1} (p+1)\, c_{u',p}. \qquad (4)$$

Part (ii) of the claim is nothing but the translation of recurrence (4) in terms of generating functions.

□

We notice here that Lemma 1 is essentially equivalent to Bentley's observation that the probability distribution of the shapes of k-d-trees constructed by n random insertions (forgetting about key values) coïncides with the corresponding distribution on 1-d trees. This probability as a function of the shape of the tree is given in [Kn73,6.2.2 ex.5].

Our next step consists in reducing the equations of Lemma 1 to a vectorial differential system of order 2k-s. The first k-components of the solution of the system represent the quantities:

$$d_u(z),\ d_{u'}(z),\ d_{u''}(z),\ldots,\ d_{u^{(k-1)}}(z) \tag{5}$$

and the remaining k-s components are the primitives of those functions in (5) whose specification pattern starts with a star (*).

Lemma 2: The function $d_u(z)$ is the first component, $y_1(z)$, of the solution of the differential system of order 2k-s:

$$\frac{d}{dz}[\underset{\sim}{y}(z)] = \Omega(z)\ \underset{\sim}{y}(z) + \underset{\sim}{b}(z) \tag{Σ}$$

where: $\underset{\sim}{y}(z) = (y_1(z),\ y_2(z),\ldots,y_{2k-s}(z))^T$

$$\underset{\sim}{b}(z) = (b_1(z),\ b_2(z),\ldots,b_{2k-s}(z))^T$$

with $b_i(z) = \dfrac{\varepsilon_i}{(1-z)^3}$: $\varepsilon_i=0$ if $i>k$; $\varepsilon_i=1$ if $i\le k$ and $u_i=S$;

$\varepsilon_i=2$ if $i\le k$ and $u_i=*$.

The initial conditions are $y_j(0)=0$. The transition matrix $\Omega(z)$ admits the block decomposition:

$$\Omega(z) = \begin{pmatrix} A & C \\ B & D \end{pmatrix}$$

where matrices A,B,C,D have respective dimensions k x k, (k-s) x k, k x (k-s), (k-s) x (k-s) and elements given by:

(i) $A_{ii}= 0$ if $u_i=S$; $A_{ii}= \dfrac{1}{z(1-z)}$ if $u_i=*$;

$A_{i,i+1 \bmod k}= \dfrac{2}{1-z}$; other elements are all zero;

(ii) $B_{ij}= 1$ if j is the rank of the i-th unspecified attribute in u, and $B_{ij}=0$ otherwise;

(iii) $C= \dfrac{1}{z^2(1-z)}\ B^T$;

(iv) D is the zero (k-s) x (k-s) matrix.

Proof: Let $\pi_1, \pi_2,\ldots\pi_{k-s}$ be the ranks of the unspecified attributes in u; ranks are assumed to be numbered from 1. For instance if u= *SS**S*, then $\pi_1= 1$, $\pi_2= 4$, $\pi_3= 5$, $\pi_4= 7$.

We set up a differential system for the quantities $y_1(z)$, $y_2(z) \ldots y_{2k-s}(z)$, where:

$$y_j(z) = d_{u^{(j-1)}}(z) \qquad \text{for } j: 1 \le j \le k \tag{6}$$

and $$y_{k+j}(z) = \int_0^z d_{u^{(\pi_j - 1)}}(t)\,dt, \quad \text{for } j: 1 \le j \le k-s \tag{7}$$

The differential relations between the y_j's are obtained as follows:

(a) if $j \le k$ and $w = u^{(j-1)}$ starts with an S, differentiating the relation given by Lemma 1 - (ii), we have:

$$d'_w(z) = \frac{2}{(1-z)^3} + \frac{2}{1-z}\, d_{w'}(z). \tag{8}$$

(b) if $j \le k$ and $w = u^{(j-1)}$ starts with a $*$, differentiating the relation given by Lemma 1-(i), we find:

$$c'_w(z) = \frac{1}{(1-z)^2} + \frac{2}{1-z}\, c_{w'}(z) \; ;$$

multiplying this relation by z and adding to both members $c_w(z)$, we find

$$d_w(z) \equiv z c'_w(z) + c_w(z) = \frac{z}{(1-z)^2} + c_w(z) + \frac{2z}{1-z}\, c_{w'}(z).$$

We now multiply this last relation by $(1-z)$, differentiate then multiply again by $\frac{1}{1-z}$ and isolate $d'_w(z)$, so that we get:

$$d'_w(z) = \frac{1}{(1-z)^3} + \frac{1}{z(1-z)}\, d_w(z) + \frac{1}{z^2(1-z)} \int_0^z d_w(z)\, dz + \frac{2}{1-z}\, d_{w'}(z) \tag{9}$$

since

$$c_w(z) = \frac{1}{z} \int_0^z d_w(z)\, dz.$$

(c) Finally relation (7) is clearly equivalent to:

$$y'_{k+j}(z) = y_{\pi_j}(z). \tag{10}$$

Putting together relations (8), (9), (10) leads to a differential system for the y_j's defined by (6), (7) and the matricial form of this system is none other than the one given by the statement of the lemma.

□

As an illustration of Lemma 2, we consider the specification pattern u= S*S, so that k = 3 and s =2 . The system is then of order 4, and its form is:

$$\frac{d}{dz}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 & \frac{2}{1-z} & 0 & 0 \\ 0 & \frac{1}{z(1-z)} & \frac{2}{1-z} & \frac{1}{z^2(1-z)} \\ \frac{2}{1-z} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} + \begin{pmatrix} \frac{2}{(1-z)^3} \\ \frac{1}{(1-z)^3} \\ \frac{2}{(1-z)^3} \\ 0 \end{pmatrix}$$

At this stage, we start plunging into complex analysis. Since by their combinatorial origin, the coefficients $c_{u,n}$ satisfy

$$c_{u,n} = 0\,(n) \tag{11}$$

we thus know that functions $c_u(z)$ and $d_u(z)$ are analytic in the domain $|z| < 1$. From the general theory of linear differential systems† in the complex plane follows that a solution to system Σ:

$$\frac{d}{dz}\,[\underset{\sim}{y}(z)] = \Omega(z)\ \ \underset{\sim}{y}(z) + \underset{\sim}{b}(z) \tag{12}$$

is analytic in any region where the coefficient matrix Ω and the vector b are analytic. The only singularities of Ω and b are at z=0 and z=1. However, we have seen that the solution defined by our initial conditions is analytic around 0. Thus there only remains z=1 as the unique singularity of the vector y(z). We propose to estimate the coefficients of $d_u(z) \equiv y_1(z)$ by means of the

† Here and in the sequel, we shall refer to the book by Henrici ([He77], chapter 9) as our main source on differential systems.

complex integral:

$$d_{u,n} = \frac{1}{2i\pi} \int_{\Gamma} d_u(z) \frac{dz}{z^{n+1}} \tag{13}$$

where Γ is any contour that simply encircles the origin inside the domain of analycity of $d_u(z)$. Following [Od82] and [FO82], we propose to choose for Γ a contour that comes close to the singularity z=1. To evaluate the integral (12) then requires detailed expansions for the solutions to system Σ around this point. The matrix $\Omega(z)$ being meromorphic with a single pole at z=1, the homogeneous system (defined by setting $\underset{\sim}{b}$ to D in (12)) has what is known as a singularity of the first kind and the y_j's are expected to have a logarithmic singularity at z=1. We shall see that the dominant contribution in the local expansion of $d_u(z)$ there is of the form

$$d_u(z) \sim \delta.(1-z)^{\lambda} \quad \text{as } z \to 1 \tag{14}$$

with

λ the smallest root of the indicial equation:

$$\det(\Omega_o - \lambda I) = 0. \tag{15}$$

where

I is the (2k-s) x (2k-s) identity matrix and

$$\Omega_o = \lim_{z \to 1} (z-1)\,\Omega(z).$$

The use of an appropriate contour Γ in (13), shows that we can translate the approximation of a function (14) into an approximation for its Taylor coefficients†

$$d_{u,n} \sim [z^n]\ \delta(1-z)^{\lambda}. \tag{16}$$

Now, the asymptotics of the coefficients of the right hand side of (16) is well known; thus , provided that δ is non zero:

---

† We let $[z^n]\ f(z)$ denote as usual the coefficient of z^n in the Taylor expansion of f(z).

$$d_{u,n} \sim \frac{\delta}{\Gamma(-\lambda)}\, n^{-\lambda-1}. \qquad (17)$$

with $\Gamma(s)$ denoting the Euler gamma function. Theorem 1 then follows from the explicit form of the indicial equation (15).
To proceed with this programme, we now prove the key proposition that describes the behaviour of function $d_u(z)$.

Proposition 1: Around z=1, the function $d_u(z)$ has an expansion of the form:

$$d_u(z) = \frac{h_{\alpha_1}}{(1-z)^{\alpha_1}} + \sum_{\alpha \in I \setminus \{\alpha_1\}} \frac{g_\alpha(\log(z-1))}{(1-z)^\alpha} + O\left(\frac{1}{(1-z)^2}\right)$$

where I is the set of all complex roots α of the equation

$$\alpha^s(\alpha-1)^{k-s} - 2^k = 0$$

satisfying $\mathrm{Re}(\alpha) \geq 2$,

$$\alpha_1 = \max(I)$$

and each $g_\alpha(u)$ is a polynomial of degree at most 5.

We shall summarize here the discussion of the proof; details can be filled in by referring to the extensive treatment given by Henrici. The general solution of the non-homogeneous system Σ is the sum of a particular solution and of the general solution of the homogeneous system :

$$\frac{d}{dz}[\underset{\sim}{w}(z)] = \Omega(z)\ \underset{\sim}{w}(z) \qquad (19)$$

We thus study separately the solutions to the homogeneous system (Lemma 3) and then construct a particular solution (Lemma 4).
There is however a difficulty that arises in this process: in differential systems logarithmic terms may be introduced when some confluences occur in expansions. As we shall see, the distribution is based on the roots of the indicial equation (15) and complications occur when two such roots differ by an integer. We need to distinguish two cases (labelled A and B) in lemmas 3, 4 depending on the following condition $\mathcal{H}_{k,s}$:

$$\mathcal{H}_{k,s} : \forall\, \lambda \neq \lambda' \ [\chi(\lambda) = 0 \text{ and } \chi(\lambda') = 0 \Rightarrow \lambda - \lambda' \notin \mathbb{Z}]$$

where the polynomial $\chi(\lambda)$ related to the indicial equation (14) is defined by

$$\chi(\lambda) = (-\lambda)^s \ (-1-\lambda)^{k-s} - 2^k \ .$$

This condition is satisfied for instance by all integers k,s: $0<s<k\leq 10$.

Lemma 3^A : If k and s satisfy condition $\mathcal{H}_{k,s}$, then, around z=1, any solution of the homogeneous system (19) has an expansion of the form:

$$\underset{\sim}{w}(z) = \sum_{\alpha\in I} \frac{\underset{\sim}{h}_\alpha}{(1-z)^\alpha} + O\left(\frac{1}{(1-z)^2}\right)$$

for some constant vectors $\underset{\sim}{h}_\alpha$.

Proof: A fundamental matrix W of system (19) is defined as a matrix whose columns form a linearly independent set of solutions, and thus it satisfies the matrix differential system:

$$\frac{d}{dz} W(z) = \Omega(z) \ W(z). \tag{20}$$

The matrix $\Omega(z)$ is meromorphic at z=1 and we can write

$$\Omega(z) = \frac{1}{z-1} \sum_{m\geq 0} \Omega_m (z-1)^m.$$

The matrix Ω_o is in the case of system Σ of the form

$$\Omega_o = - \left(\begin{array}{c|c} A_o & C_o \\ \hline 0 & 0 \end{array}\right)$$

with A_o a matrix of dimension k x k whose elements are found from matrix A:

$A_{o,ii} = 0$ if $u_i = S$; $A_{o,ii} = 1$ if $u_i = *$; $A_{o,i,i+1 \bmod k} = 2$;

other elements are all equal to 0. The matrix C_o is equal to B^T (B defined in Lemma 2).

Returning to our previous example where u=S*S, we have for instance:

$$\Omega_o = \begin{pmatrix} 0 & -2 & 0 & 0 \\ 0 & -1 & -2 & -1 \\ -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The characteristic polynomial of matrix Ω_o is determined by successive expansions along the last k-s rows:

$$\text{char}\,(\Omega_o) = (-\lambda)^{k-s}\ \text{char}(A_o)$$

and a direct calculation from the definition of A_o shows that

$$(-1)^k.\ \text{char}(A_o) = (-\lambda)^s\ (-1-\lambda)^{k-s} - 2^k. \tag{21}$$

This is the polynomial $\chi(\lambda)$ introduced in the definition of $\mathcal{H}_{k,s}$. Since

$$\chi'(\lambda) = (-1)^k \lambda^{s-1}\ (1+\lambda)^{k-s-1}\ (k\lambda+s)$$

we directly check that $\chi(\lambda)$ has only simple roots for $s \neq 0,k$. Thus A_o can be diagonalized and its block structure shows that the same property holds true for Ω_o. Therefore for some transition matrix T, we have:

$$\Omega_o = T^{-1}\ \Delta\ T$$

where Δ is diagonal with last k-s diagonal elements equal to zero. The system

$$W'(z) = \frac{\Omega_o}{z-1}\ W \tag{22}$$

can be viewed as an "approximation" to system (20); it has a fundamental matrix of the form

$$W(z) = (z-1)^{\Delta}. \tag{23}$$

Considering the system (20) as a perturbation of system (22), one proves by the method of indeterminate coefficients that (20) has a solution of the form

$$W(z) = P(z)\ \ (z-1)^{\Delta} \tag{24}$$

where P is analytic at z=1 and $P(1)=T^{-1}$ under the condition that no two roots of $\chi(\lambda)$ differ by an integer.

We have assumed here that k and s satisfy this condition $\mathcal{H}_{k,s}$. Expressed differently equation (24) then means that every component W_j of a solution $\underset{\sim}{W}$ of the homogeneous system (19) has a finite expansion of the form

$$W_j(z) = \sum_{\alpha} \frac{h_{\alpha}^{(j)}(z)}{(1-z)^{\alpha}} + h_o^{(j)}(z) \tag{25}$$

for some functions $h_\alpha^{(j)}(z)$ analytic at z=1 where the sum is over α's solution of the equation

$$\chi(-\alpha) = 0 \quad \text{or} \quad \alpha^s(\alpha-1)^{k-s} - 2^k = 0 \tag{26}$$

(with again $\chi(\lambda)$ defined by (21)). The term $h_o(z)$ corresponds to the eigenvalue 0 of matrix Ω_o.

To conclude with the proof of Lemma 3^A, we therefore only need to study the localisation of the exponents α in equation (26). Since these are zeros of the polynomial

$$\chi(-\alpha) = \alpha^s(\alpha-1)^{k-s} - 2^k$$

they have to satisfy

$$|\alpha| < 3$$

for all values of k and s and there is always a unique zero α_1 of $\chi(-\alpha)$ in the interval (2,3). Furthermore, it is easy to check that all other roots of $\chi(-\alpha)$ have a real part strictly less than α_1. (Actually it can be also proved that when k is large enough $\chi(-\alpha)$ has several complex roots whose real parts are in the interval $(2,\alpha_1)$). We thus obtain the statement of the lemma by selecting in (25) only those terms whose α satisties $\mathrm{Re}(\alpha)>2$ and retaining only the first terms $h_\alpha^{(j)}(1)$ of the $h_\alpha^{(j)}(z)$. □

Lemma 3^B: If k and s do not satisfy condition $\mathcal{H}_{k,s}$, then around z=1 any solution to the homogeneous system (19) has an expansion of the form

$$\underset{\sim}{W}(z) = \frac{\underset{\sim}{h}_{\alpha_1}}{(1-z)^{\alpha_1}} + \sum_{\alpha\in I\setminus\{\alpha_1\}} \frac{\underset{\sim}{g}_\alpha(\log(z-1))}{(1-z)^\alpha} + 0\left(\frac{1}{(1-z)^2}\right),$$

where each component $g_\alpha^{(j)}(u)$ of $g_\alpha(u)$ is a polynomial of degree at most 5 in u.

Proof: The reduction method (Theorem 9.5.d of [He77], p 122) transforms a system

$$\underset{\sim}{W}' = \frac{1}{(z-1)}\,\Omega(z)\,\underset{\sim}{W}$$

where matrix $\Omega(1)$ has eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_m$ into a system

$$\hat{\underset{\sim}{W}} = \frac{1}{(z-1)}\,\hat{\Omega}(z)\,\hat{\underset{\sim}{W}} \tag{27}$$

where matrix $\hat{\underset{\sim}{\Omega}}(1)$ has eigenvalues λ_1-1, $\lambda_2,\ldots,\lambda_m$; the relation between $\underset{\sim}{W}$ and $\hat{\underset{\sim}{W}}$ being of the form

$$\underset{\sim}{W}(z) = \underset{\sim}{H}(z)\ \hat{\underset{\sim}{W}}(z) \tag{28}$$

for some analytic matrix $\underset{\sim}{H}(z)$.

Using it repeatedly, we transform the original system into a system of the form (27) with the correspondence given by (28), in such a manner that: (a) the eigenvalues of $\hat{\underset{\sim}{\Omega}}(1)$ are a subset of the eigenvalues of $\underset{\sim}{\Omega}(1)$; (b) no two eigenvalues of $\hat{\underset{\sim}{\Omega}}(1)$ differ by an integer. Furthermore, the "dominant" eigenvalue α_1 still has multiplicity 1; each other non-zero eigenvalue has multiplicity at most 6, since any root of $\chi(\lambda)$ has to satisfy $|\lambda| < 3$; finally eigenvalue 0 admits a set of k-s linearly independent eigenvectors.

A fundamental matrix of system (27) can thus be put under the form:

$$\hat{\underset{\sim}{W}}(z) = \hat{\underset{\sim}{P}}(z)\ (z-1)^{\hat{S}} \tag{29}$$

with $\hat{\underset{\sim}{P}}(z)$ analytic at 1 and $\hat{S}$ upper triangular.

However we are no longer guaranteed that $\hat{S}$ may be diagonalized. Matrix $\hat{S}$ decomposes into

$$\hat{S} = \hat{\Delta} + \hat{U}$$

with $\hat{\Delta}$ diagonal, $\hat{U}$ a strict upper triangular matrix (i.e. with all its diagonal elements being 0) which commutes with $\hat{\Delta}$ (see [He 77] p 120) and $\hat{U}^6 = 0$ (the null matrix). Thus

$$(z-1)^{\hat{S}} = e^{\hat{S}\log(z-1)} = (z-1)^{\hat{\Delta}}\ (I + \sum_{k=1}^{5} \frac{\hat{U}^k}{k!}\ (\log(z-1))^k) \tag{30}$$

Grouping (28),(29) and (30) establishes the claim of the Lemma. □

The next stage now consists in constructing a particular solution of the non-homogeneous system Σ. This is achieved by means of the matricial "variation-of-constants" formula.

Lemma 4^A: If hypothesis $\mathcal{H}_{k,s}$ is satisfied, the non-homogeneous system Σ admits in a neighbourhood of z=1 a particular solution of the form

$$\frac{\underset{\sim}{H}(z)}{(1-z)^2} + \underset{\sim}{G}(z)\ \log(z-1)$$

where $\underset{\sim}{H}(z)$ and $\underset{\sim}{G}(z)$ are analytic at z=1.

Proof: By the variation of-constants formula ([He77], p 99), if $\hat{z}$ is a regular point of the system(for instance $\hat{z} = \frac{1}{2}$) and W(z) a fundamental matrix of the homogeneous system, the general solution to the non-homogeneous system is given by

$$\underset{\sim}{W}(z) = W(z)\ W^{-1}(\hat{z})\ \underset{\sim}{c} + W(z) \int_{\hat{z}}^{z} W^{-1}(t)\ \underset{\sim}{b}(t)\ dt, \tag{31}$$

and the second term is a particular solution of the non-homogeneous system. We know that the homogeneous system has a fundamental matrix of the form

$$W(z) = P(z)\ (z-1)^{\Delta}$$

where P(z) is analytic at 1:

$$P(z) = \sum_{m\geq 0} P_m (z-1)^m$$

and P_o is regular. Thus

$$W^{-1}(z) = (z-1)^{-\Delta}\ Q(z)$$

where

$$Q(z) = \sum_{m\geq 0} Q_m\ (z-1)^m$$

and again Q_o regular. Since

$$b(z) = \frac{\underset{\sim}{b_o}}{(z-1)^3}$$

for some constant vector $\underset{\sim}{b_o}$, we find for the particular solution

$$\underset{\sim}{U}(z) = W(z) \int_{\hat{z}}^{z} W^{-1}(t)\ \underset{\sim}{b}(t)\ dt$$

the expansion:

$$\underset{\sim}{U}(z) = P(z) \left[\sum_{n\geq 0} (z-1)^{\Delta} \int_{z}^{z} (t-1)^{-\Delta-n-3}\ dt\ Q_m\ \underset{\sim}{b_o} \right]. \tag{32}$$

Integration of the matrix shows that (taking for instance $\hat{z} = \frac{1}{2}$)

$$\underset{\sim}{U}(z) = P(z) \left[\sum_{n\geq 0} (z-1)^{\Delta} \left(I_n(z) - I_n(\hat{z})\right) Q_n \underset{\sim}{b}_o \right]$$

where $I_n(z)$ is a diagonal matrix whose elements are:

- if $n\neq 2$

$$\frac{(z-1)^{-\lambda_j+n-2}}{-\lambda_j+n-2} \qquad 1\leq j\leq k$$

$$\frac{(z-1)^{n-2}}{n-2} \qquad k<j\leq 2k-s$$

- if $n=2$

$$\frac{(z-1)^{-\lambda_j}}{-\lambda_j} \qquad 1\leq j\leq k$$

$$\log(z-1) \qquad k<j\leq 2k-s. \qquad (33)$$

with $\lambda_1, \lambda_2, \ldots, \lambda_k$ the roots of polynomial $\chi(\lambda)$.

Splitting the sum in (32) we find

$$\underset{\sim}{U}(z) = \underset{\sim}{U}_1(z) - \underset{\sim}{U}_2(z)$$

where

$$\underset{\sim}{U}_1(z) = P(z) \sum_{n\geq 0} (z-1)^{\Delta} I_n(z) Q_n \underset{\sim}{b}_o$$

$$\underset{\sim}{U}_2(z) = P(z) \sum_{n\geq o} (z-1)^{\Delta} I_n(\hat{z}) Q_n \underset{\sim}{b}_o$$

The vector $\underset{\sim}{U}_2(z)$ is a solution of the homogeneous system, so that a particular solution of system Σ is provided by $\underset{\sim}{U}_1(z)$. Separating the terms in the sum according to $n\neq 2$, $n=2$, we have:

$$\underset{\sim}{U}_1(z) = P(z) \sum_{n\neq 2} (z-1)^{\Delta} I_n(z) Q_n \underset{\sim}{b}_o$$

$$+ \quad P(z) (z-1)^{\Delta} I_2(z) Q_2 \underset{\sim}{b}_o. \tag{34}$$

The diagonal form of $I_n(z)$ in (33) shows that terms of the form $(z-1)^{\pm\lambda}$ disappear in the products of (34) and we are left with

$$U_1(z) = \frac{\underset{\sim}{H}(z)}{(z-1)^2} + \underset{\sim}{G}(z) \log(z-1)$$

for some vectors $\underset{\sim}{H}$ and $\underset{\sim}{G}$ analytic at $z=1$. □

Lemma 4^B: If hypothesis $\mathcal{H}_{k,s}$ is not satisfied, then the non-homogeneous system Σ admits in a neighbounhood of $z=1$ a particular solution of the form

$$\frac{\underset{\sim}{H}(z)}{(1-z)^2} + \sum_{k=1}^{5} \underset{\sim}{G}_k(z) (\log(z-1))^k$$

where $\underset{\sim}{H}(z)$ and the $\underset{\sim}{G}_k(z)$ are analytic at $z=1$.

Proof: The previous method applied to this case would rather trivially imply the existence of a particular solution with dominant terms of the form

$$\frac{(\log(z-1))^5}{(z-1)^2}.$$

However the stronger property of the statement of the lemma is required for the later part of the analysis. It is derived by what looks like a "failed attempt" at a direct solution of system Σ by the method of indeterminate coefficients.

Let $\underset{\sim}{H}(z)$ have the expansion

$$\underset{\sim}{H}(z) = \sum_{m\geq 0} \underset{\sim}{H}_m (z-1)^m.$$

If we try to identify coefficients of $\underset{\sim}{H}(z)$ so that

$$\frac{\underset{\sim}{H}(z)}{(z-1)^2}$$

satisfies system Σ we find the equations:

$$\begin{aligned} (\Omega_o + 2I)\ \underset{\sim o}{H} &= \underset{\sim o}{b} \\ (\Omega_o + I)\ \underset{\sim 1}{H} &= -\Omega_1\ H_o. \end{aligned} \tag{35}$$

where $\underset{\sim o}{b}$ is a constant vector defined by

$$b(z) = \frac{\underset{\sim o}{b}}{(z-1)^3}\ .$$

System (35) is solvable since (Ω_o+2I) and (Ω_o+I) are non singular. The next equation would be

$$\Omega_o\ \underset{\sim 2}{H} = -\Omega_1\ \underset{\sim 1}{H} - \Omega_2\ \underset{\sim o}{H}$$

which need not be solvable since Ω_o is singular. However if $\underset{\sim}{w}(z)$ is a solution to system Σ and $\underset{\sim o}{H}$, $\underset{\sim 1}{H}$ are defined by (35), we find that

$$\underset{\sim}{\bar{w}}(z) = \underset{\sim}{w}(z) - \frac{\underset{\sim o}{H}}{(z-1)^2} - \frac{\underset{\sim 1}{H}}{(z-1)} \tag{36}$$

satisfies the modified system:

$$\frac{d}{dz}\ \underset{\sim}{\bar{w}}(z) = \Omega(z)\ \underset{\sim}{\bar{w}}(z) + \bar{b}(z) \tag{37}$$

where $\bar{b}(z)$ has now only a simple pole at z=1. It is to the transformed system (37) that we now apply the method of variation of constants. By the developments of Lemma 3^B, a fundamental matrix of the homogeneous system corresponding to (37) is of the form

$$W(z) = P(z)\ (z-1)^{\hat{R}} = P(z)\ (z-1)^{\hat{\Delta}}\ (I + \sum_{k=1}^{5} \frac{\hat{U}^k}{K!}\ (\log(z-1))^k)$$

and its inverse may be similarly written

$$W^{-1}(z) = [\, I + \sum_{k=1}^{5} (-1)^k \frac{\hat{U}^k}{k!} (\log(z-1))^k] \; (z-1)^{-\hat{\Delta}} \, Q(z).$$

We can use, as in Lemma 4^A, this form into the variation-of-constants formula. A particular solution to (37) is thus given by

$$\bar{w}(z) = P(z) \; (z-1)^{\hat{\Delta}} \, X \sum_{n \geq 0} \int_{\hat{z}}^{z} Y \; (t-1)^{-\hat{\Delta}-I} dt \; Q \, \underset{\sim}{\bar{b}}_n$$

for some vector of constants $\underset{\sim}{\bar{b}}_o$, and some matrices X,Y whose coefficients are polynomial in log(z-1); X and Y also commute with $\hat{\Delta}$ or with matrices of a similar block structure like $(z-1)^{\pm\Delta}$. Carrying out the integration explicitly leads to

$$\bar{w}(z) = \underset{\sim}{\bar{H}}(z) + \sum_{k=1}^{5} \underset{\sim}{G}_k(z) \; \log(z-1)^k$$

which, combined with (36) yields the claim of the lemma. □

We can now conclude with the proof of Proposition 2: the most general solution to system Σ is obtained as a sum of the particular solution $\underset{\sim}{v}(z)$ constructed in Lemma 5 which satisfies

$$W(z) = O\left(\frac{1}{(z-1)^2}\right)$$

and of the general solution to the homogeneous system whose behaviour is described in Lemma 3.

The next stage consists in translating the expansion of $d_u(z)$ around its singularity z=1 into information about the asymptotics of its coefficients. This uses the following result:

Proposition 2: (i) The n-th Taylor coefficient c_n of the function

$$c(z) = (1-z)^{-\alpha} [\log(1-z)]^k$$

satisfies:

$$c_n = n^{\alpha-1} \, \Pi(\log n) + O(n^{\alpha-2} \log^k n)$$

for some polynomial Π (depending on α and k) of degree at most k.

(ii) Suppose that g(z) is analytic in

$$E = \{z : |z| \leq 1,\ z \neq 1\}$$

and that for $z \in E$

$$g(z) = O(|1-z|^{-\beta})$$

for some $\beta > 0$. Then the n-th Taylor coefficient g_n of g(z) satisfies:

$$g_n = O(n^{\beta-1}).$$

Proof: Part (ii) of the proposition is taken from [FO82] (Proposition 7 p 209): it is proved there by expressing c_n by means of the Cauchy integral formula and taking as a contour of integration the circle of convergence of g(z) except for a small notch inside the circle at distance $\frac{1}{n}$ of the singularity 1. (See also [082] for related results). Proof of part (i) of the proposition relies on similar methods, except that now precise asymptotic results are needed. One starts from the integral form of c_n:

$$c_n = \frac{1}{2i\pi} \int_{\Gamma} c(z)\ \frac{dz}{z^{n+1}} \tag{38}$$

and use for Γ the contour (oriented anticlockwise):

$$\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$$

where:

$$\Gamma_1 = \{z = 1 + \frac{e^{i\theta}}{n} : \theta \in [\frac{\pi}{2} ; \frac{3\pi}{2}]\}$$

$$\Gamma_2 = \{z = 1 + \frac{i}{n} + \frac{x}{n} : x \in [1;n]\}$$

$$\Gamma_3 = \{z : |z| = (4 + \frac{1}{n^2})^{1/2},\ |Re(z)| < 2\}$$

$$\Gamma_4 = \{z : \bar{z} \in \Gamma_2\}.$$

This contour is depicted in Figure 2. Decomposing the integral (38) along the particular contour Γ, we have:

$$c_n = c_n^{(1)} + c_n^{(2)} + c_n^{(3)} + c_n^{(4)}.$$

Since $c(z)$ is bounded along Γ_3:

$$c_n^{(3)} = O(2^{-n}), \tag{39}$$

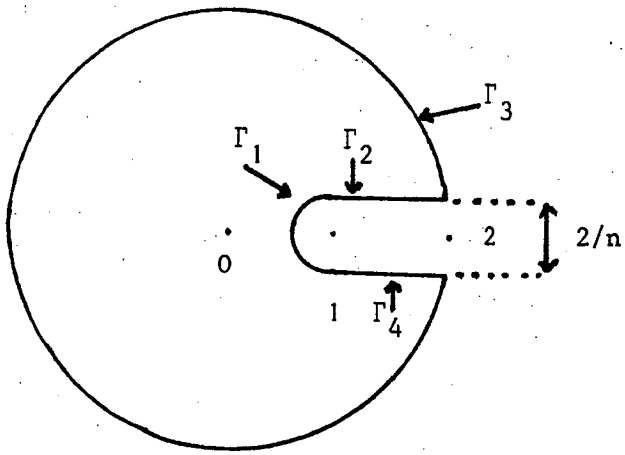


Figure 2: The contour Γ used in the proof of Proposition 2.

so that we only have to evaluate $c_n^{(1)}$ on the one hand and $c_n^{(2)}$, $c_n^{(4)}$ on the other hand. Proceeding with the evaluation of $c_n^{(1)}$, the change of variable:

$$z = 1 + \frac{e^{i\theta}}{n}$$

shows that:

$$c_n^{(1)} = -\frac{n^{\alpha-1}}{2\pi} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} (i\theta - \log n)^k \, e^{-i(\alpha-1)\theta} \, (1 - \frac{e^{i\theta}}{n})^{-(n+1)} \, d\theta$$

Using the exponential approximation

$$(1 - \frac{e^{i\theta}}{n})^{-(n+1)} = e^{e^{i\theta}} + O(\tfrac{1}{n})$$

in the previous integral, we find:

$$c_n^{(1)} = -\frac{n^{\alpha-1}}{2\pi} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} (i\theta - \log n)^k \; e^{i(\alpha-1)\theta} \; e^{e^{i\theta}} d\theta + O((\log n)^k \; n^{\alpha-2}). \qquad (40)$$

The integral there is clearly a polynomial of degree at most k in logn. For the integral along Γ_2, we use the change of variable:

$$z = 1 + \frac{i}{n} + \frac{x}{n}$$

with which we find:

$$c_n^{(2)} = \frac{n^{\alpha-1}}{2i\pi} \int_0^n \frac{(\log(-i-x)-\log n)^k}{(-i-x)^\alpha} \; (1 + \frac{x+i}{n})^{-(n+1)} \; dx.$$

Similarly using the exponential approximation for $x \leq n^{1/3}$

$$(1+ \frac{x+i}{n})^{-(n+1)} = e^{-i-x} \; (1+0(\frac{x+x^2}{n}))$$

in $c_n^{(2)}$ (the integral from $n^{1/3}$ to n is exponentially small), we find:

$$c_n^{(2)} = \frac{n^{\alpha-1}}{2i\pi} \int_0^{n^{1/3}} \frac{(\log(-i-x)-\log n)^k}{(-i-x)^\alpha} \; e^{-i-x} \; dx + O((\log n)^k \; n^{\alpha-2}).$$

The integral can be extended from 0 to ∞ introducing only exponentially small terms, so that

$$c_n^{(2)} = n^{\alpha-1} \; \frac{e^i}{2i\pi} \int_0^\infty \frac{(\log(-i-x)-\log n)^k}{(-i-x)^\alpha} \; e^{-x} \, dx + O((\log n)^k \; n^{\alpha-2}) \qquad (41)$$

the integral being again a polynomial in log n of degree at most k. The case of $c_n^{(4)}$ is entirely similar, and combining (39), (40), (41) establishes the claim of the Lemma . Notice also that the same method would make it possible to determine a complete asymptotic expansion of c_n for any fixed k and fixed k and n. Finally part (ii) of the proposition shows that the polynomials Π cannot be identically zero. □

Now a direct application of Proposition 2 to the result of Proposition 1 shows that $d_{u,n}$ admits the asymptotic expansion:

$$d_{u,n} = \frac{h_{\alpha_1}}{\Gamma(\alpha_1)} n^{\alpha_1 - 1} + \sum_{\alpha \in I \setminus \{\alpha_1\}} \xi_\alpha(\log n)\, n^{\alpha - 1} + O(n)$$

where the $\xi_\alpha(u)$ are polynomials of degree at most 5. Since

$$c_{u,n} = \frac{1}{n+1} d_{u,n},$$

we thus have:

$$c_{u,n} = \frac{h_{\alpha_1}}{\Gamma(\alpha_1)} n^{\alpha_1 - 2} + \sum_{\alpha \in I \setminus \{\alpha_1\}} \xi_\alpha(\log n)\, n^{\alpha - 2} + O(1). \qquad (42)$$

To complete the proof of Theorem 1, we therefore only need to show that the actual order of $c_{u,n}$ is given by the first term of (42):

$$c_{u,n} = \Omega(n^{\alpha_1 - 2})$$

or equivalently:

$$h_{\alpha_1} \neq 0.$$

Lemma 5: The coefficient h_{α_1} in the expansion of $c_{u,n}$ is strictly positive.

Proof: The proof which is non-constructive proceeds through an indirect argument using the positivity of the $c_{u,n}$ and a logarithmic lowerbound on the $c_{u,n}$. Assume a contrario that $h_{\alpha_1} = 0$.

(i) if all the h_α were equal to zero, then we would have $c_{u,n} = O(1)$, $d_{u,n} = O(n)$ as $n \to \infty$.

This contradicts the fact that $c_{u,n}$ is at least as large as the cost of a completely specified search which is known to be $O(\log n)$. The analytic equivalent of this argument consists in observing that

$$d_{u,n} \geq d_{\sigma,n} \quad \text{where } \sigma = SS\ldots S,\ |\sigma| = k,$$

as follows from recurrences (2), (4). Then the solution of the equation for $d_\sigma(z)$:

$$d'_\sigma(z) = \frac{2}{(1-z)^3} + \frac{2}{1-z}\, d_\sigma(z)$$

is found to be

$$d_\sigma(z) = \frac{2}{(1-z)^2}\, \log\left(\frac{1}{1-z}\right)$$

so that

$$d_{\sigma,n} \sim 2\, n \log n$$

whence a contradiction in this case.

(ii) Thus if $h_{\alpha_1}=0$, atleast one of the h_α for $\alpha\in I\setminus\{\alpha_1\}$ is non zero. The complex roots of the indicial equation

$$\chi(-\alpha) = 0$$

occur in pairs of complex conjugates. Let therefore β, $\bar{\beta}$ be the roots of highest real part such that

$$h_\beta \neq 0\ ,\ h_{\bar{\beta}} = \bar{h}_\beta \neq 0;$$

from Proposition 2 follows that for some constant $C \neq 0$:

$$c_{u,n} \sim C n^{\beta-2} (\log n)^k + \bar{C} n^{\bar{\beta}-2} (\log n)^k$$

for some k : $0\leq k\leq 5$. Thus with

$$C = a+ib\ ,\quad \beta=\sigma+it$$

we find:

$$c_{u,n} \sim 2n^{\sigma-2}\ (\log n)^k\ (a\cos(t\log n)-b\sin(t\log n)).$$

But such an equation contradicts the fact that the $c_{u,n}$ are non-negative numbers.

We have thus seen that in all cases the assumption $h_{\alpha_1}=0$ leads to a contradiction, so that Lemma 5 is established.

This lemma itself allows us to complete the proof of Theorem 1: the dominant exponent e in the asymptotic form of $c_{u,n}$:

$$c_{u,n} \sim \gamma_u \, n^e$$

is $e=\alpha_1-2$, and it is the unique positive real root of the equation

$$\chi(-2-e) = 0$$

that is

$$(e+2)^s \; (e+1)^{k-s} - 2^k = 0$$

or equivalently

$$(e+2)^{s/k} \; (e+1)^{1-s/k} - 2 = 0,$$

and the function $\theta(\frac{s}{k})$ is

$$\theta(\frac{s}{k}) = e-(1-\frac{s}{k})$$

which therefore satisfies the equation of the statement of Theorem 1.

3- DIGITAL TECHNIQUES FOR INTERNAL AND EXTERNAL SEARCH

In this section, we provide an analysis of partial match retrieval for k-d-tries (in Section 3.1) and for grid file algorithms (Section 3.2). Our basic interest is in the so-called Bernoulli model corresponding to the description given in Section 1 : the number of keys in the file is a fixed integer n and keys are assumed to be taken independently from a uniform distribution. As a consequence of these hypotheses, bits of arbitrary positions in arbitrary fields of keys are independent uniform {0,1} random variables. There is also strong interest in a closely related model, called the Poisson model (see for instance [FNPS79] for analyses under this model) : there, the number of keys in the file is assumed to be a random variable N with a Poisson distribution, i.e. such that

$$\Pr(N=k) = e^{-n} \frac{n^k}{k!}$$

for some fixed parameter n which corresponds to the expectation of N. The interest of the Poisson model is to make sometimes technical developments

simpler because of some strong independence properties of the localization of keys in non-overlapping subintervals.

We have analyzed k-d-tries and grid files under both the Bernoulli and Poisson models, and the main orders of costs appear to be idential. For the sake of conciseness, we illustrate the analytic techniques involved by giving only the proof of the evaluation of k-d-tries under the Bernoulli model and of grid-file algorithms under the Poisson model.

3.1- Multidimensional tries

Consider again a file $F \subset D_1 x D_2 x \ldots x D_k$ where each attribute domain D_i is assimilated to the set of infinite binary sequences:

$$D_i \tilde{=} \{0,1\}^{\infty}.$$

To any record $r=(r_1,r_2,\ldots r_k)$ is associated an infinite binary sequence in the usual manner through regular shuffling:

let

$$r_j = r_j^{(1)}\ r_j^{(2)}\ r_j^{(3)}\ \ldots\ ,\ r_j^{(k)} \in \{0,1\}$$

be the binary representation of attribute r_j; the infinite sequence associated to r is

$$\rho = \text{shuffle}(r) = \rho^{(1)}\ \rho^{(2)}\ \rho^{(3)} \ldots\ ,\ \rho^{(k)} \in \{0,1\}$$

where

$$\rho = r_1^{(1)},\ r_2^{(1)},\ \ldots,\ r_k^{(1)},\ r_1^{(2)},\ r_2^{(2)} \ldots,\ r_k^{(2)},\ r_3^{(1)},\ r_2^{(3)},\ldots,r_k^{(3)},\ldots.$$

Thus the shuffle of a k-tuple is obtained by taking in sequence the first bit of attribute 1, the first bit of attribute 2, ..., the first bit of attribute k, then starting cyclically again with the second bits of attributes 1,2,...k, etc....

By definition, the k-d-trie constructed on a finite set F is the (1-d-) trie constructed on the set $\{\text{shuffle}(r)/r \in F\}$. Thus k-d-tries have some analogy to k-d-trees with the notable difference that the partitioning of elements corresponds to fixed values of the fields instead of values provided by the file itself, and records are stored at the leaves of the tree. The fact that 1-d-tries tend to be better balanced than 1-d-search trees does not crucially affect the performances of one dimensional search which are logarithmic in both cases. However, in the context of multidimensional search it leads to asymptotically smaller orders as we now prove it.

Theorem 2: The average cost of a partial match query of specification pattern u with s specified attributes in a k-d-trie constructed from a file of either size n (under the Bernoulli model) or expected size n (under the Poisson model) satisfies

$$c_{u,n} = \gamma(\frac{1}{k}\log_2 n)\ n^{1-s/k} + O(1)$$

where $\gamma(u)$ is a periodic function of u with period 1, small amplitude and mean value

$$\gamma_o = -\frac{s}{k^2 \log 2}\ \Gamma(\frac{s}{k} - 1)\ \sum_{\ell=0}^{k-1} (\delta_1\ \delta_2 \ldots \delta_1)\ 2^{-\ell(1-s/k)}$$

with $\delta_\ell = 1$ if the ℓ-th attribute of the query is specified, and $\delta_\ell = 2$ if it is unspecified.

As announced earlier, we only give here the proof of the estimate under the Bernoulli model. The proof under the Poisson model follows trivially by adapting the methods introduced in Section 3.2 below.

Lemma 6: The exponential generating function of the average costs $c_{u,n}$ under the Bernoulli model:

$$c_u(z) = \sum_{n\geq o} c_{u,n}\ \frac{z^n}{n!}$$

satisfies the relation:

$$c_u(z) = \delta_1 e^{z/2}\ c_u'(z/2) + e^z - 1 - z$$

with u' obtained by circularly shifting the letters of u by one position to the left.

Proof: Let $t = t_1 \hat{} t_2$. be a k-d-trie associated to a particular file F. If the first attribute of the query is non specified, we have, for the expected cost of a random query:

$$c_u[t] = 1 + c_u'[t_1] + c_u'[t_2] \qquad (1)$$

since the search then has to proceed in parallel along both subtrees with

attributes changing cyclically according to pattern u'. If the first attribute is specified, on the contrary, we find

$$c_u[t] = 1+\frac{1}{2}\,(c_{u'}[t_1] + c_{u'}[t_2]) \qquad (2)$$

since with probability $\frac{1}{2}$ the first bit of the first attribute of the query starts with a 0 (the search then proceeds in t_1) and with probability $\frac{1}{2}$ it starts with a 1 (the search then proceeds in t_2).

Given n random elements $n\geq 2$ organized in a k-d-trie $t=t_1\hat{}t_2$, the probability that

$$|t_1| = p \qquad |t_2| = n-p$$

is given by the Bernoulli probabilities:

$$\binom{n}{p}\left(\frac{1}{2}\right)^{p}\left(\frac{1}{2}\right)^{n-p} = \frac{1}{2^n}\binom{n}{p}.$$

whence for the expected values the recurrences

$$\text{if } u=*v \qquad c_{u,n} = 1+\frac{2}{2^n}\sum_p \binom{n}{p} c'_{u,p} \qquad n\geq 2;$$

$$\text{if } u=Sv \qquad c_{u,n} = 1+\frac{1}{2^n}\sum_p \binom{n}{p} c'_{u',p} \qquad n\geq 2.$$

In general, for all u and n we therefore have

$$c_{u,n} = 1+\frac{\delta_1}{2^n}\sum_p \binom{n}{p} c'_{u,p} - \delta_{n,o} - \delta_{n,1}. \qquad (3)$$

The translation of (3) in terms of exponential generating functions yields the claim of the lemma. □

Lemma 7: The generating function $c_u(z)$ satisfies the difference equation

$$c_u(z) = 2^{k-s} e^{z(1-1/2^k)} c_u\left(\frac{z}{2^k}\right) + \sum_{j=o}^{k-1} (\delta_1\delta_2\ldots\delta_j)\, e^{z(1-1/2^j)}\cdot\left(e^{z/2^j} - 1 - \frac{z}{2^j}\right)$$

Proof: From Lemma 6, we see that $c_u(z)$ is the first component of a vectorial system of difference equations:

$$(\Sigma)\qquad \begin{cases} c_u(z) = \delta_1 e^{z/2}\, c_u'(z/2) + e^z-1-z \\ c_u'(z) = \delta_2 e^{z/2}\, c_u''(z/2) + e^z-1-z \\ \cdots \\ c_u^{(k-1)}(z) = \delta_k e^{z/2}\, c_u(z/2) + e^z-1-z \end{cases}$$

This system can be solved by successive eliminations. Let $a(z)$ denote e^z-1-z. Transporting the expression of $c_u'(z)$ given by the second equation inside the defining equation for $c_u(z)$, we find

$$c_u(z) = a(z) + \delta_1 e^{z/2}\, a(z/2) + \delta_1\delta_2\, e^{z/2}\, e^{z/4}\, c_u''(z/4).$$

We continue in this fashion, using the equation satisfied by c_u'', c_u''', until the relation is only in terms of $c_u(z)$ itself. □

A functional equation of the form satisfied by $c_u(z)$, namely:

$$\phi(z) = \alpha e^{\beta z}\, \phi(\gamma z) + A(z)$$

(with ϕ the unknown function) may be solved formally by iteration in a manner similar to the proof of Lemma 7:

$$\begin{aligned} \phi(z) &= A(z) + \alpha e^{\beta z} A(\gamma z) + \alpha^2\, e^{(\beta+\beta\gamma)z}\, \phi(\gamma^2 z) \\ &= A(z) + \alpha e^{\beta z} A(\gamma z) + \alpha^2\, e^{(\beta+\beta\gamma)z}\, A(\gamma^2 z) + \alpha^3 e^{(\beta+\beta\gamma+\beta\gamma^2)z}\, \phi(\gamma^3 z) \\ &\quad \cdots\cdot \\ &= \sum_{j\geq o}^{\infty} \alpha^j \exp\left(\beta\left(\frac{1-\gamma^{j+1}}{1-\gamma}\right)z\right) A(\gamma^j z). \end{aligned}$$

Thus using here the particular form of α, β, γ and $A(z)$ we find

$$c_u(z) = \sum_{j=o}^{\infty} 2^{j(s-k)} \left\{ \left[e^z - e^{z(1-1/2^{kj})} \left(1 + \frac{z}{2^{kj}}\right)\right] + \right.$$

$$\left. \delta_1 \left[e^z - e^{z(1-1/2\,2^{kj})} \left(1 + \frac{z}{2.2^{kj}}\right)\right] + \delta_1\delta_2 \left[e^z - e^{z(1-1/4\,2^{kj})} \cdot \left(1 + \frac{z}{4.2^{kj}}\right)\right] + \cdots \right\}.$$

where inside the infinite summation we have a sum of k terms.
Extracting the Taylor coefficients of $c_u(z)$ given by this sum, we get:

Lemma 8: The expected cost of a partial match retrieval has for $n \geq 2$ the explicit form:

$$c_{u,n} = \sum_{\ell=0}^{k-1} \delta_1 \delta_2 \ldots \delta_\ell \sum_{j \geq 0} 2^{j(k-s)} \tau_{j,\ell}(n) \qquad (4)$$

where for j and ℓ not both zero:

$$\tau_{j,\ell}(x) = 1 - (1-2^{-kj-\ell})^x - x2^{-kj-1}(1-2^{-kj-\ell})^{x-1} \qquad (5)$$

and

$$\tau_{0,0}(x) = 1.$$

We observe that the convergence of (4) is guaranteed by the fact that, for fixed n, as j tends to infinity:

$$\tau_{j,\ell}(n) \sim 1 - \exp(-n2^{-kj-\ell}) - n2^{-kj-\ell}\exp(-n2^{-kj-\ell}) = O(n^2 2^{-2k}). \qquad (6)$$

Indeed the exponential approximation (6) is usually the starting point of asymptotic evaluations, but here we shall use a different approach (see [Re83] for other applications) which is more direct and may be used to obtain asymptotic expansions to any order if required.

We also observe that each $\tau_{j,\ell}$ is a positive number at most 1, so that if we sum on j=1 to ∞ in (4), we introduce an error term that is bounded above by $k.2^{s-k}$:

$$c_{u,n} = \phi(n) + O(1)$$

where $$\phi(x) = \sum_{\ell=0}^{k-1} \delta_1 \delta_2 \ldots \delta_\ell \sum_{j \geq 1} 2^{j(k-s)} \tau_{j,\ell}(x) \qquad (7)$$

Equations (5) and (7) thus define $\phi(x)$ for arbitrary real $x \geq 0$. We propose to perform the asymptotic analysis of $\phi(x)$ by investigating properties of its

Mellin transform given by

$$\phi^*(\sigma) = \int_0^\infty \phi(x)\, x^{\sigma-1}\, dx. \tag{8}$$

It is known (see for instance [Do55], [Da78]) that under suitable analytic condition, the asymptotic properties of $\phi(x)$ as $x\to\infty$ are directly related to the singularities of $\phi^*(s)$ in a right half plane. We therefore need to derive an expression for $\phi^*(s)$ that reveals some of its singularities and provides an analytic continuation of the integral definition (8). We prove:

Proposition 3: The Mellin transform of the function $\phi(x)$ given by equation (7) and such that:

$$c_{u,n} = \phi(n) + O(1)$$

has the form:

$$\phi^*(\sigma) = -(1+\sigma)\Gamma(\sigma)\left[\frac{2^{k(\sigma-\sigma_o)}}{1-2^{k(\sigma-\sigma_o)}} + A(\sigma)\right] \sum_{\ell=0}^{k-1} \delta_1 \delta_2 \dots \delta_\ell\, 2^{\ell\sigma}$$

where $\sigma_o = -(1-\frac{s}{k})$ and $A(\sigma)$ is analytic in $-1 \le \mathrm{Re}(\sigma) \le \frac{s}{2k}$ and satisfies in this region

$$A(\sigma) = O(|\sigma|^2).$$

Proof: We appeal to the following classical properties of Mellin transforms:

(i)
$$\int_0^\infty (e^{-x}-1)\, x^{\sigma-1} dx = \Gamma(\sigma); \qquad -1<\mathrm{Re}(s)<0$$

(ii)
$$\int_0^\infty (xe^{-x})\, x^{\sigma-1} dx = \sigma\,\Gamma(\sigma); \qquad -1<\mathrm{Re}(s)$$

(iii)
$$\int_0^\infty f(ax) x^{\sigma-1}\, dx = a^{-\sigma} \int_0^\infty f(x)\, x^{\sigma-1}\, dx; \qquad a>0.$$

Writing $\tau_{j\ell}(x)$ under the form:

$$\tau_{j\ell}(x)=1 - \exp(-x\alpha_{j\ell}) - \beta_{j\ell}^{x} \exp(-x\alpha_{j\ell})$$

with

$$\alpha_{j\ell} = - \log(1-2^{-kj-\ell})$$

$$\beta_{j\ell} = 2^{-kj-\ell} (1-2^{-kj-\ell})^{-1}$$

we find thus that the Mellin transform of $\tau_{j\ell}(x)$ is:

$$\tau^{*}_{j\ell} = -(\alpha_{j\ell})^{-\sigma} \Gamma(\sigma) - \beta_{j\ell} (\alpha_{j\ell})^{-s-1} \sigma\Gamma(\sigma) \qquad (9)$$

provided $-1<\mathrm{Re}(\sigma) < 0$. From (9), we can determine the expression of $\phi^{*}(\sigma)$ applying the linearity of the transform to the defining equation (7). The conditions on the values of σ, in order for the interchange of integration in (8) and the infinite summation in (7) to be justified, are that the sums

$$\omega_{\ell}(\sigma)= \sum_{j\geq 1} 2^{j(k-s)} (\alpha_{j\ell})^{-\sigma}; \ \omega'_{\ell}(\sigma)= \sum_{j\geq 1} 2^{j(k-s)} \beta_{j\ell} (\alpha_{j\ell})^{-\sigma-1} \qquad (10)$$

be absolutely convergent. Using the asymptotic equivalents

$$(\alpha_{j\ell})^{-\sigma} = O(2^{kj\mathrm{Re}(\sigma)}), \ \beta_{j\ell}=O(2^{-kj})$$

we see that the sums defining $\omega_1(\sigma)$ and $\omega'_1(\sigma)$ are uniformly and absolutely convergent when σ is in any stripe:

$$S_{\eta}: \ -1<\mathrm{Re}(\sigma) < -(1- \tfrac{s}{K}) -\eta \ , \ \eta>0. \qquad (11)$$

Thus the transform of $\phi(x)$ is defined in the S_o strip and there:

$$\phi^{*}(\sigma) = - \sum_{\ell=o}^{k-1} \delta_1 \delta_2 \ldots \delta_{\ell} (\omega_{\ell}(\sigma) + \sigma\omega'_{\ell}(\sigma)). \ \Gamma(\sigma) \qquad (12)$$

the next stage consists in analytically continuing $\phi^{*}(\sigma)$,

that is to say of the $\omega_\ell(\sigma)$ to a domain that extends to the right of $\sigma_o = -(1-\frac{s}{k})$. To that purpose we use the expansion valid for small u, uniformly in σ for σ in any fixed stripe $c<Re(\sigma)<d$:

$$(-\log(1-u))^{-\sigma} = u^{-\sigma}(1-\frac{\sigma u}{2} + O(|\sigma|^2 u^2)). \tag{13}$$

this expansion suggests "approximating" $\omega_\ell(\sigma)$ and $\omega'_\ell(\sigma)$ by the series:

$$\hat{\omega}_\ell(\sigma) = \sum_{j\geq 1} 2^{j(k-s)} \left(2^{kj+\ell}\right)^{\sigma}$$

This series can be summed exactly when $Re(\sigma) < \sigma_o = -(1-\frac{s}{k})$:

$$\hat{\omega}_\ell(\sigma) = 2^{\ell\sigma} \frac{2^{k-s+k\sigma}}{1-2^{k-s+k\sigma}} \tag{14}$$

and expansion (11) shows that the differences $\omega_\ell(\sigma) - \hat{\omega}_\ell(\sigma)$ and $\omega'_\ell(\sigma) - \hat{\omega}_\ell(\sigma)$ have a general term of the form

$$O(2^{j(kRe(\sigma)-s)})$$

and therefore are analytic for $Re(\sigma) < \frac{s}{k}$. Equation (13) also shows that

$$\omega_\ell(\sigma) - \hat{\omega}_\ell(\sigma) = O(|\sigma|^2) \quad ; \quad \omega'_\ell(\sigma) - \hat{\omega}_\ell(\sigma) = O(|\sigma|^2) \tag{15}$$

for large $|\sigma|$ with $-1 \leq Re(\sigma) \leq \frac{s}{2k}$.

Thus

$$\phi^*(\sigma) = -\Gamma(\sigma)\ (1+\sigma) \sum_{\ell=0}^{k-1} \delta_1\delta_2\ldots\delta_\ell\ \hat{\omega}_\ell(\sigma)$$

$$-\Gamma(\sigma) \sum_{\ell=0}^{k-1} \delta_1\delta_2\ldots\delta_\ell [\omega_\ell(\sigma) - \hat{\omega}_\ell(\sigma) + \sigma(\omega'_\ell(\sigma) - \hat{\omega}_\ell(\sigma))]$$

and using (14), (15) concludes the proof of the proposition. □

The final stage to conclude with the asymptotic analysis of $\phi(x)$ for large x, thus with the asymptotics of $c_{u,n}$, is to use the inversion theorem for Mellin transforms:

$$\phi(x) = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} \phi^*(\sigma) x^{-\sigma} d\sigma \ , \quad -1<c<- \ (1- \frac{s}{k}) \tag{16}$$

and, under suitable conditions evaluate the integral using Cauchy's theorem as a sum of residues to the right of the vertical line $\{c+it \ / t \in \mathbb{R}\}$ and a remainder term of a small order when x is large.

We consider the integral

$$\phi_N(x) = \frac{1}{2i\pi} \int_{\Gamma_N} \phi^*(\sigma) x^{-\sigma} d\sigma \tag{17}$$

where Γ_N is the rectangular contour oriented clockwise (see Figure 3)

$$\Gamma_N = \Gamma_N^1 + \Gamma_N^2 + \Gamma_N^3 + \Gamma_N^4 \tag{18}$$

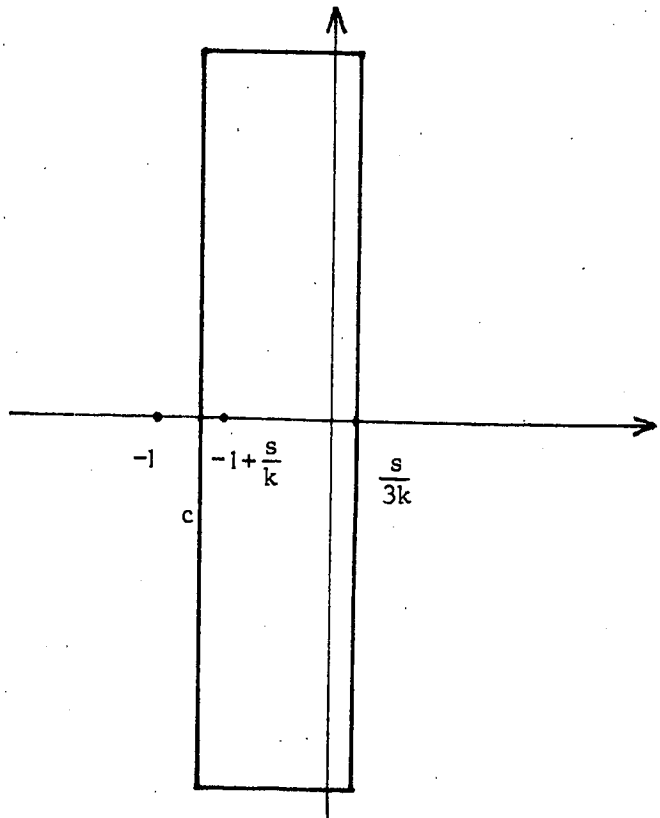


Figure 3: The rectangular contour Γ_N used in evaluating $\phi(x)$ through an inverse Mellin transform.

$$\Gamma_N^1 = \{c+it \ : \ |t| \le \frac{(2N+1)\pi}{k\log 2} \}$$

$$\Gamma_N^2 = \{u+ \frac{(2N+1)i\pi}{k\log 2} \ : \ c \le u \le \frac{s}{3k} \}$$

$$\Gamma_N^3 = \{ \frac{s}{3k} + it \ : \ |t| \le \frac{(2N+1)\pi}{k\log 2} \}$$

$$\Gamma_N^4 = \{u- \frac{(2N+1)i\pi}{k\log 2} \ : \ c \le u \le \frac{s}{3k} \},$$

with N an integer (contours of a similar type are used for instance in [Kn73] p132).
Setting

$$\phi_N(x) = \phi_N^1(x) + \phi_N^2(x) + \phi_N^{(3)} + \phi_N^{(4)}$$

where ϕ_N^j corresponds to the contributions to integral (17) of the part Γ_N^j of the contour, we have the following results:

(i) $\phi_N^1(x) \to \phi(x)$ as $N \to \infty$

(ii) $\phi_N^2(x) = o(1)$ as $N \to \infty$

(iii) $|\phi_N^3(x)| \le x^{-\frac{s}{3k}} \int_{\Gamma_\infty} |\phi^*(\sigma)| \, d\sigma = O(x^{-\frac{s}{3k}})$

(iv) $\phi_N^4(x) = o(1)$ as $N \to \infty$.

Of these assertions (i) is obvious by continuity; (ii) and (iv) come from the exponential decrease of $\Gamma(s)$ towards $i\infty$; (iii) is the trivial majorization of the absolute value of an integral.

Thus letting N tend to infinity, we find:

$$\phi_\infty(x) = \phi(x) + O(x^{-\frac{s}{3k}}). \qquad (18)$$

Now the integral (17) can also be evaluated as the sum of the residues of the integrand inside Γ_N. As $N=\infty$, this sum is absolutely convergent and we have:

$$\phi_\infty(x) = - \sum_{\alpha \in \text{Pole}(\phi^*(\sigma))} \text{Res}(\phi^*(\sigma)x^{-\sigma}, \sigma=\alpha) \tag{19}$$

The poles of $\phi^*(\sigma)$ inside Γ_∞ are:

- simple poles at

$$\alpha_j = \sigma_o + \frac{2ij\pi}{k\log 2}$$

- a simple pole at $\sigma=0$.

Thus (18) (19) can be rewritten as:

$$\phi(x) = - \sum_{\alpha \in \text{Pole}(\phi^*(\sigma))} x^{-\alpha} \text{Res}(\phi^*(\sigma), \sigma=\alpha) + O(x^{-\frac{s}{3k}})$$

which is precisely an asymptotic expansion of $\phi(x)$ for large x. The contribution of the pole $\alpha=0$ is $O(1)$; the contribution of $\alpha=\sigma_o$ is derived from the result of Proposition 2, and is:

$$x^{-\sigma_o} \frac{(1+\sigma_o)\ \Gamma(\sigma_o)}{k\log 2} \sum_{\ell=0}^{k-1} \delta_1\delta_2 \cdots \delta_\ell\ 2^{\ell\sigma_o}. \tag{20}$$

The contribution of α_j is similarly:

$$x^{-\sigma_o}\ e^{-\frac{2ij\pi}{k}\log_2 x}\ (1+\alpha_j)\ \Gamma(\alpha_j) \sum_{\ell=0}^{k-1} \delta_1\delta_2 \cdots \delta_\ell 2^{\ell\alpha_j} \tag{21}$$

so that

$$c_{u,n} = n^{1-s/k}\ \gamma\left(\frac{\log_2 x}{k}\right)$$

with $\gamma(u)$ a periodic function of u with period 1, mean value and Fourier coefficients obtained from (20), (21) respectively.

3.2- Grid file algorithms

Grid file or extendible cell methods are a class of algorithms suitable for maintaining large collections of multiattribute records on secondary storage (see [NHS81],[LR82], [Ta82]).
They are based on a dynamically varying partitioning of the underlying record space that adapts itself gracefully to the particular structure of the file being operated on. These algorithms can be viewed as multidimensional generalization of Dynamic Hashing [La78], Extendible Hashing [FNPS79] or Virtual Hashing [Li78].

If a suitable splitting policy is used (like in [Ta82] or [NHS81] when one uses level alternation for attribute splittings instead of time alternation), the paging of the file is equivalent to the paging of a k-d-trie.

Definition: The paged k-d-trie with page capacity b built on a file F is obtained from the k-d-tree built on F by placing in single pages all maximal subtrees containing at most b records.

The part of the tree obtained by pruning all leaf pages is called the index (or directory) of the paged k-d-trie.

This definition is illustrated by Figure 4 .

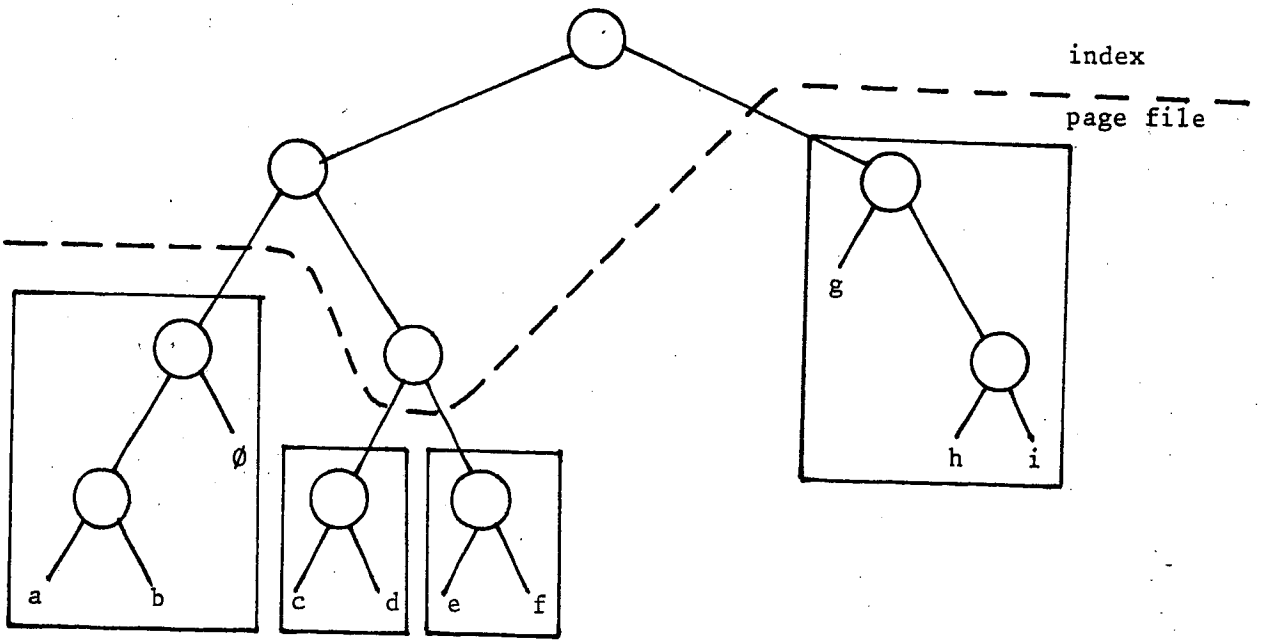


Figure 4: The paged 2-d-trie corresponding to F={a,b,...,i} with page capacity 3 when: a=(00-,00-); b=(00-,01-); c=(00-,10-); d=(00-,11-); e=(01-,10); f=(01-,11-); g=(1-,0-); h=(10-,1-); i=(11-,1-).

The various schemes mentioned above differ by the way the index is implemented: it may be kept in core (like in Dynamic Hashing of [La78]), it may be represented as a perfect tree embedding encoded into an array ([Ta82] generalizing [FNPS79]) or as a multidimensional array [NHS81].

The characteristic parameter of the cost of a partial match retrieval that is independent of the particular representation of the index is the number of accesses to the paged file. Its expected value is given by the following theorem.

Theorem 3: The expected cost of a partial match query measured by the number of page accesses in a paged k-d-trie constructed from a file of size n (under the Bernoulli model) or from a file of expected size n (under the Poisson model) satisfies:

$$c_{u,n} = \gamma\left(\frac{1}{k}\log_2 n\right) n^{1-s/k} + O(1)$$

where $\gamma(u)$ is a periodic function of u with period 1 and mean value

$$\gamma_o = \frac{\Gamma(\frac{s}{k}-1)\,(\frac{s}{k}+b-1)}{k\log 2}\,[(\delta_1-1)+\delta_1(\delta_2-1)2^{\frac{s}{k}-1}+\ldots+\delta_1\delta_2\ldots\delta_{k-1}(\delta_k-1)\,2^{(k-1)(\frac{s}{k}-1)}],$$

The expected cost measured in the number of page accesses corresponding to a tree $t = t_1 \hat{} t_2$ for a random query with specification pattern u satisfies:

$$c_u[t] = \begin{cases} \frac{\delta_1}{2}\,(c_u'[t_1]+c_u'[t_2]) & \text{if } |t|>b \\ 1 & \text{if } |t|\le b. \end{cases}$$

It proves necessary for our later treatment to operate with the modified quantity

$$\hat{c}_u[t] = c_u[t] - 1;$$

it satisfies for all t the recurrence:

$$c_u[t] = \frac{\delta_1}{2}\,(c_u'[t_1] + c_u'[t]) + (\delta_1 - 1)\,\chi(|t| > b) \tag{22}$$

with $\chi(P)$ the characteristic function of predicate P.
The expected value $\hat{c}_{u,n} = c_{u,n} - 1$ of $\hat{c}_u[t]$ taken over all trees of size n under the Bernoulli model therefore satisfies:

$$\hat{c}_{u,n} = \delta_1 \sum_{k=0}^{n} \frac{1}{2^n} \binom{n}{k} \hat{c}_{u,k}' + (\delta_1 - 1)\,\chi(n > b) \tag{23}$$

The exponential generating function of the $c_{u,n}$:

$$\hat{c}_u(z) = \sum_{n \geq 0} \hat{c}_{u,n} \frac{z^n}{n!}$$

satisfies a relation obtained from (23):

$$\hat{c}_u(z) = \delta_1\, e^{z/2}\, \hat{c}_u'(z/2) + (\delta_1 - 1)\,[e^z - e_b(z)] \tag{24}$$

with $e_b(z)$ denoting the truncated exponential:

$$e_b(z) = \sum_{j=0}^{b} \frac{z^j}{j!}\,. \tag{25}$$

We now let $d_u(z)$ denote the quantity:

$$d_u(z) = e^{-z}\, \hat{c}_u(z).$$

Thus $d_u(z)$ is the expectation of $\hat{c}_u[t]$ if the number of elements in the file follows a Poisson distribution with parameter z.
Equation (24) then leads to a difference system relating $d_u(z)$, $d_u'(z)$,...:

$$\left\{\begin{array}{l} d_u(z) = \delta_1\, d_u'(z/2) + (\delta_1-1)\ (1-e^{-z}e_b(z)) \\ d_u'(z) = \delta_2\, d_u''(z/2) + (\delta_2-1)\ (1-e^{-z}e_b(z)) \\ \qquad\qquad \cdots \\ d_u{}^{(k-1)}(z) = \delta_k\, d_u(z/2) + (\delta_k-1)\ (1-e^{-z}e_b(z)) \end{array}\right. \qquad (\Sigma)$$

From the combinatorial origin of parameters, we know that $d_u(z)=0(z^{b+1})$ for small z and $d_u(z)=0(z)$ for large z.
Thus the Mellin transforms of $d_u(z)$, $d_u'(z)$... are all defined in the stripe

$$-(b+1) < \mathrm{Re}(\sigma) < -1. \qquad (26)$$

We let now $d^*_u(\sigma)$ denote the Mellin transform of $d_u(z)$. From functional properties of the Mellin transforms recalled in the previous section follows that the transforms satisfy the linear system:

$$\left\{\begin{array}{l} d^*_u(\sigma) = \delta_1 2^{\sigma}\, d^*_{u'}(\sigma) + (\delta_1-1)\ \alpha(\sigma) \\ d_u'(\sigma) = \delta_2 2^{\sigma}\, d^*_{u''}(\sigma) + (\delta_1-1)\ \alpha(\sigma) \\ \qquad\qquad \cdots \\ d^*_{u(k-1)}(\sigma) = \delta_{k-1} 2^{\sigma}\, d^*_u(\sigma) + (\delta_1-1)\ \alpha(\sigma) \end{array}\right. \qquad (\Sigma^*)$$

where

$$\alpha(\sigma) = \int_0^{\infty} (1-e_b(x)e^{-x})\ x^{\sigma-1}\ dx.$$

This last transform is also defined in the stripe (26), and it can be computed by linearity. We find:

$$\alpha(\sigma) = -\sum_{j=0}^{b} \frac{\Gamma(\sigma+j)}{j!} = -\Gamma(\sigma)\ \beta_b(\sigma)$$

where

$$\beta_b(\sigma) = 1 + \frac{\sigma}{1!} + \frac{\sigma(\sigma+1)}{2!} + \ldots + \frac{\sigma(\sigma+1)\ldots(\sigma+b-1)}{b!} = \binom{\sigma+b}{b} . \qquad (27)$$

The system Σ^* can be solved in a manner similar to what was done before for k-d-tries; successively eliminating $d^*_{u'}$, $d^*_{u''}$... we get:

$$d^*_u(\sigma) = 2^{k-s}\, 2^{k\sigma}\, d^*_u(\sigma) - \beta_b(\sigma)\, \Gamma(\sigma)\, \omega(\sigma)$$

Whence by solving for $d^*_u(\sigma)$:

$$d^*_u(\sigma) = \frac{-\beta_b(\sigma)\, \Gamma(\sigma)\omega(\sigma)}{1-2^{k(\sigma-\sigma_o)}} \qquad (28)$$

with:

$$\omega(\sigma) = (\delta_1 - 1) + \delta_1(\delta_2 - 1)2^{\sigma} + \delta_1\delta_2(\delta_3 - 1)2^{2\sigma} + \ldots + \delta_1\delta_2\ldots\delta_{k-1}(\delta_k - 1)2^{(k-1)\sigma} \qquad (29)$$

and

$$\sigma_o = -\left(1 - \frac{s}{k}\right).$$

We can now conclude with the asymptotic analysis recovering $d_u(z)$ from $d^*_u(\sigma)$ by means of the inversion theorem for Mellin transforms, using the contours Γ_N of section 3.2 and calculating residues as was done before.

□

Acknowledgements: Some of this work was started while the first author was visiting the Tata Institute of Fundamental Research in Bombay. This author would like to express his gratitude to Pr. Narasimhan, M. Joseph, R.K. Shyamasundar for their invitation and to S. Joshi for several discussions that lead to this work.

REFERENCES

[Be75] Bentley J.L.: Multidimensional binary search trees used for associative searching.
CACM, 18, 9 (1975), pp 509-517.

[CS77] Cardenas A.F., Sagamang J.P.: Doubly-chined tree data base organization - Analysis and design strategies.
Compt. J., 20, (1977), pp 15-26.

[Da78] Davies: Integral transforms and their applications.
Springer-Verlag (1978).

[Do55] Doetsch: Handbuch der Laplace transformation, band II: Anwendungen der Laplace transformation, 1. Abteilung.
Birkhaueser Verlag, Basel und Stuttgart, (1955).

[FNPS79] Fagin R., Nievergelt J., Pippenger N., Strong H.R.: Extendible Hashing - A fast access method for dynamic files.
ACM TODS, 4,3, (1979), pp 315-344.

[FB74] Finkel R.A., Bentley J.L.: Quad trees: A data structure for retrieval on composite keys.
Acta Informatica, 4, (1974), pp 1-9.

[FO82] Flajolet P., Odlyzko A.: The average height of binary trees and other simple trees.
JCSS, 25, 2, (1982), pp 171-213.

[FP83] Flajolet P., Puech C.: Tree structures for partial match retrieval.
Rapport interne LRI n°128, Université de Paris-Sud (1983).

[He77] Henrici P.: Applied and complex computational analysis, vol. 2.
Wiley, New York, (1977).

[KSY77] Kashyap R.L., Subas S.K.C., Yao S. Bing: Analysis of the multiple-attribute-tree data-base organization.
IEEE Transactions on Software Engineering, SE-3, 6, (1977), pp 451-467.

[Kn73] Knuth D.E.: The art of computer programming, vol. 3.
Addison-Wesley, (1973).

[La78] Larson P.A.: Dynamic hashing.
BIT, 18, (1978), pp 184-201.

[Li78] Litwin W.: Virtual hashing: a dynamically changing hashing.
Proc. 4th Conf. on VLDB, Berlin, (1978), pp 517-522.

[LR82] Lloyd J.W., Ramamohanarao K.: Partial-match retrieval for dynamic files.
BIT, 22, (1982), pp 150-168.

[NHS81] Nievergelt J., Hinterberger H., Sevcik K.C.: The grid file: an adaptable, symmetric multi-key file structure.
ETH Report, 46, (1981).

[Re83] Regnier M.: Evaluation des performances du hachage dynamique.
Thèse de 3ème cycle, Université de Paris-Sud, (Avril 1983).

[Ri76] Rivest R.L.: Partial-match retrieval algorithms.
Siam J. Comput., 5, 1, (1976), pp 19-50.

[Ta82] Tamminen M.: The extendible cell method for closest point problems.
BIT, 22, (1982), pp 27-41.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique