23

Arbres digitaux et équations aux différences

Philippe Flajolet
INRIA, Rocquencourt

[résumé par Jean-Marc Steyaert]

Les arbres digitaux constituent une structure de données très riche qui apparaît dans plusieurs contextes d'application : arbres de recherche, stratégies de hachage dynamique, recherche multi-dimensionnelle, protocoles de réseaux, algorithmes probabilistes, etc. L'analyse de ces structures et algorithmes met en évidence trois modèles principaux que sous-tend la même structure d'arbre binaire :

I. les *tries* pour lesquels les sommets internes sont vides et les feuilles seules contiennent une clef ;

II. les *arbres de recherche digitaux* pour lesquels tous les sommets, internes et externes, contiennent une clef ;

III. les *arbres paginés* pour lesquels les sommets contiennent b clefs pour les internes et jusqu'à b clefs pour les feuilles.

La distribution du nombre de clefs dans les sous-arbres est dans les trois cas de type Bernoulli équilibré : la probabilité que le sous-arbre gauche contienne p clefs et le sous-arbre droit q clefs valant $\binom{p+q}{p}2^{-(p+q)}$. Les deux paramètres d'intérêt sont le nombre de sommets internes S_n(qui décrit la compacité) de l'arbre et la longueur de cheminement L_n(qui décrit le degré d'équilibrage) de l'arbre, dont on calcule les valeurs moyennes en fonction du nombre n de clefs contenues dans l'arbre. Pour les trois modèles on attend des comportements respectifs du type $s_n = E(S_n) = O(n)$ et $l_n = E(L_n) = O(n \log n)$.
Bien que les résultats dus respectivement à Knuth-de Bruijn [1], Flajolet-Sedgewick [2] et Flajolet-Richmond [3] soient semblables, les méthodes mises en œuvre vont en se compliquant comme le laisse suggérer la forme des équations (différentielles) aux différences que satisfont les s.g.e. de dénombrement des divers paramètres :

$$\begin{aligned} \text{I.} \quad & f(z) &=& \ a(z) + 2e^{z/2} f(z/2), \\ \text{II.} \quad & \tfrac{d}{dz} f(z) &=& \ a(z) + 2e^{z/2} f(z/2), \\ \text{III.} \quad & \tfrac{d^k}{dz^k} f(z) &=& \ a(z) + 2e^{z/2} f(z/2). \end{aligned}$$

La forme de ces équations est bien sûr intimement liée au processus de Bernoulli. Montrons à titre d'exemple la formule de type I pour la taille des *tries*. Si f_n est la valeur moyenne de la taille, elle satisfait la récurrence : $f_n = 1 - \delta_{n,0} + \sum_{p+q=n} \binom{n}{p} 2^{-p} 2^{-q} (f_p + f_q)$, où $\delta_{n,0}$ est le symbole de Kronecker. Ces récurrences se résolvent classiquement en passant aux s.g.e. et on obtient pour la s.g.e. $f(z) = \sum_{n \geq 0} f_n z^n / n!$ l'équation $f(z) = e^z - 1 + 2e^{z/2} f(z/2)$. Lorsqu'il l'on place une clef dans les sommets internes la sommation se fait pour $p + q = n - 1$, et il est bien connu que pour "compenser" il faut dériver ce qui donne le type II, et semblablement pour le type III.

1 Les *tries*

La résolution algébrique des équations de type I se fait par itération et on obtient en toute généralité : $f(z) = \sum_{k\geq 0} 2^k a(z/2^k)\exp(z(1-2^{-k}))$. Si $a(z)$ a une forme simple, l'extraction des coefficients de la série est routinière.

Théorème 1 [Knuth] : *Les valeurs moyennes s_n de la taille et l_n de la longueur de cheminement des* tries *ont les expressions exactes :*

$$\begin{aligned} s_n &= \sum_{k\geq 0} 2^k(1-(1-2^{-k})^n - \frac{n}{2^k}(1-2^{-k})^{n-1}), \\ l_n &= n\sum_{k\geq 0}(1-(1-2^{-k})^{n-1}). \end{aligned}$$

L'asymptotique de ces sommes peut se faire de manière élémentaire si l'on se contente de l'ordre de grandeur. On observe en effet que les termes valent essentiellement 1 pour k petit et 0 pour k grand, avec une transition relativement brutale pour $k \sim \log_2 n$. On est déduit alors simplement : $s_n = O(n)$ et $l_n = n\log_2 n + O(n)$.
De façon plus fine, en utilisant une technologie classique en théorie analytique des nombres, Knuth et de Bruijn ont mis en évidence le caractère fluctuant de ses quantités autour d'une valeur moyenne dont l'ordre de grandeur a été obtenu précédemment.

Théorème 2 [Knuth et de Bruijn] : *Les valeurs moyennes s_n de la taille et l_n de la longueur de cheminement des* tries *ont les expressions asymptotiques :*

$$\begin{aligned} 1+2s_n &= \frac{2n}{\log 2} + nQ(\log_2 n) + O(\sqrt{n}), \\ l_n &= n\log_2 n + nP(\log_2 n) + O(\sqrt{n}), \end{aligned}$$

où P et Q sont deux fonctions périodiques de période 1.

La méthode requise pour évaluer les sommes telles que s_n et l_n, que nous appellerons *harmoniques*, repose sur la transformation intégrale de Mellin :

$$f^\star(s) = \int_0^\infty f(x)x^{s-1}dx,$$

qui associe donc à une fonction f ayant de bonnes propriétés de croissance à l'origine ($f(x) = O(x^\alpha)$) et à l'infini ($f(x) = O(x^\beta)$), la fonction $f^\star(s)$ qui est analytique dans la bande $-\alpha < \Re(s) < -\beta$. La formule inverse de Perron

$$f(x) = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty} f^\star(s)x^{-s}ds, \text{ pour } -\alpha < c < -\beta,$$

permet souvent après étude des singularités de $f^\star$ d'en déduire le comportement asymptotique de $f(x)$, quand x tend vers l'infini.
Illustrons l'ensemble du processus dans le cas de la longueur de cheminement L_n d'un arbre de taille n qui vérifie comme variable aléatoire dépendant de x, nombre de feuilles du sous-arbre gauche, la récurrence :

$$\begin{aligned} L_n &= n + L_x + L_{n-x}, \\ L_0 &= L_1 = 0. \end{aligned}$$

Comme $\mathbf{P}(x = k|n \text{ clefs}) = \pi_{nk} = 2^{-n}\binom{n}{k}$, on en déduit que l'espérance de L_n satisfait la récurrence :

$$l_n = n - \delta_{n1} + \sum_{k=0}^{n} 2^{-n} \binom{n}{k} (l_k + l_{n-k}).$$

La résolution d'une telle récurrence se fait de manière (maintenant) classique en utilisant les s.g.e., $f_n \mapsto f(z) = \sum f_n z^n/n!$, et les constructions admissibles :

$$\begin{aligned} 1 &\mapsto e^z, \\ n &\mapsto ze^z, \\ \delta_{n0} &\mapsto 1, \\ \delta_{n1} &\mapsto z, \\ \frac{f_n}{2^n} &\mapsto f(z/2), \\ \sum_{k=0}^{n} \binom{n}{k} f_k &\mapsto e^z f(z). \end{aligned}$$

Dans ces conditions, il est aisé de déduire que la s.g.e. $l(z)$ des l_n vérifie l'équation fonctionnelle :

$$l(z) = z(e^z - 1) + 2e^{z/2} l(z/2),$$

qui se résout par itération en

$$l(z) = \sum_{k\geq 0} 2^k e^{z(1-\frac{1}{2^k})} \frac{z}{2^k} (e^{z(1-\frac{1}{2^k})} - 1).$$

Après simplification, l'extraction des coefficients de Taylor de $l(z)$ est un jeu d'enfant et fournit l'expression du théorème 1.
Passons maintenant à l'étude du comportement asymptotique de l_n. Un calcul d'approximation un peu fastidieux, fondé sur le fait que quand k est supérieur à $\log_2 n$, $(1 - 2^{-k})^n \approx e^{-n/2^k}$, permet de remplacer l'expression exacte de l_n/n par la somme harmonique approchée suivante :

$$\frac{l_n}{n} = \sum_{k\geq 0} (1 - e^{-(n-1)/2^k}) + o(1).$$

Le problème revient donc à estimer le comportement asymptotique de la fonction $\Lambda(x) = \sum_{k\geq 0}(1 - e^{-x/2^k})$, lorsque x tend vers l'infini.
Remarquons d'abord que la transformée de Mellin de e^{-x} n'est autre que la bien connue fonction Gamma d'Euler : $\Gamma(s) = \int_0^\infty e^{-x} x^{s-1} dx$. La transformée de Mellin $\lambda^\star(s)$ de $\lambda(x) = 1 - e^{-x}$, est également $\Gamma(s)$, mais prise dans la bande $-1 < \Re(s) < 0$ (c 'est un des moyens de prolonger analytiquement $\Gamma(s)$). Par changement de variable trivial et linéarité, on a alors :

$$\Lambda^\star(s) = -\sum_{k\geq 0} 2^{ks} \Gamma(s) = \frac{-\Gamma(s)}{1 - 2^s},$$

fonction qui doit être considérée dans la bande fondamentale $-1 < \Re(s) < 0$. Le développement asymptotique pour $\Lambda(x)$, quand x tend vers l'infini, est obtenu en appliquant à $\Lambda^\star(s)$ la formule de

Perron sus-mentionnée, dans la bande fondamentale , puis pour calculer cette intégrale en déplaçant l'abscisse d'intégration vers la droite tout en collectant les contributions de chacun des pôles de $\Lambda^\star(s)$ rencontrés au moyen de la formule des résidus : on utilise ici le fait que l'intégrande tend vers 0 lorsque la partie imaginaire de s tend vers l'infini. Le fait que $\Lambda^\star(s)$ possède un pôle double en 0 (celui de $\Gamma(s)$ et de $(1-2^s)^{-1}$) et des pôles simples en $s = \frac{\pm 2ik\pi}{\log 2}$, $k > 0$, permet de conclure que :

$$\Lambda(x) = \frac{\log x}{\log 2} + \frac{1}{2} + \frac{\gamma}{\log 2} - \sum_{k\neq 0} \frac{\Gamma(\frac{2ik\pi}{\log 2})}{\log 2} x^{-\frac{2ik\pi}{\log 2}},$$

la dernière somme mettant en évidence la périodicité de la fonction P du théorème 2 (et sa décomposition en série de Fourier), qui s'obtient immédiatement.

2 Les arbres digitaux

Ces arbres ont été étudiés par Knuth, Konheim et Newman pour ce qui est de leur longueur de cheminement et par Flajolet et Sedgewick pour ce qui est du nombre de sommets internes-externes, c'est-à-dire en frange de l'arbre.

Théorème 3 : *Les valeurs moyennes de la longueur de cheminement interne et du nombre de sommets internes-externes sont respectivement :*

$$(n+1)\log n + \frac{\gamma - 1}{\log 2} + \frac{1}{2} - \alpha + W(\log_2 n) + O(\sqrt{n})$$

et

$$\lambda.n + Q(\log_2 n).n + O(\sqrt{n}),$$

où $\alpha = 1 + 1/3 + 1/7 + 1/15 + 1/31 + ...$ *et* $W(u)$ *et* $Q(u)$ *sont périodiques de période 1.*

De fait $Q(u)$ a la forme surprenante suivante, $Q(u) = (1-u/2)(1-u/4)(1-u/8)...$
Comme indiqué dans l'introduction, les quantités ci-dessus se définissent récursivement sur les sous-arbres gauche et droit, et dans ce cas le fait qu'une clef soit rangée dans chaque sommet interne (il y en a donc n) induit des récurrences légèrement différentes, du type :

$$f_n = a_n + \sum_{k=0}^{n} \pi_{nk}(f_k + f_{n-1-k}),$$

avec $\pi_{nk} = 2^{-(n-1)}\binom{n-1}{k}$, puisque seules $n-1$ clefs restent à placer.
Un calcul algébrique classique conduit ainsi aux équations différentielles aux différences de type II. Le schéma de résolution consiste à passer par les séries génératrices de Poisson : $g(z) = e^{-z}f(z) = \sum f_n e^{-z} z^n/n! = \sum_n g_n z^n/n!$. L'équation générique de type II se transforme alors en :

$$g'(z) + g(z) = \alpha(z) + 2g(z/2).$$

Ainsi a-t-on pour les coefficients la récurrence :

$$g_{n+1} + g_n = \alpha_n + 2^{1-n} g_n,$$

qui dans le cas de la longueur de cheminement, $a(z) = ze^z$, $\alpha(z) = z$, conduit à la formule explicite $g_n = (-1)^n \prod_{j=1}^{n-2}(1-2^{-j})$, puis par inversion à la forme explicite pour l'espérance de la longueur de cheminement :

$$E(l_n) = \sum_{k=2}^{n} \binom{n}{k} (-1)^k Q_{k-2},$$

avec $Q_m = (1-1/2)(1-1/4)(1-1/8)\ldots(1-1/2^m)$.
On est donc en présence d'une différence dont il faut faire l'asymptotique. La première remarque est que pour la plus part des fonctions, u^m, $\sqrt{u}$, $\log u$, $1/(u^2+1)$, $1/(1-2^u)$, etc. ces différences (qui s'apparentent à des dérivées de très grand ordre) sont faibles.
La clef de leur asymptotique est fournie par les intégrales de Rice :

$$S_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k f(k) = \frac{(-1)^n}{2i\pi} \int_{\mathcal{C}} f(z) \frac{n!}{z(z-1)(z-2)\ldots(z-n)} dz,$$

où $\mathcal{C}$ est un contour qui entoure simplement les points $0, 1, 2, \ldots, n$.
Si de plus f ne croît pas trop vite à l'infini dans le demi-plan droit, il est possible de remplacer l'intégrale de contour par une sommation le long d'une droite imaginaire

$$(-1)^n \Delta^n f = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} f(z) \frac{n!}{z(z-1)(z-2)\ldots(z-n)} dz,$$

puis si f est méromorphe d'appliquer la technique déjà vue dans le cas précédent de balayer les pôles, vers la gauche cette fois, en collectant les résidus :

$$(-1)^n S_n = \sum Res f(z) \frac{n!}{z(z-1)(z-2)\ldots(z-n)} dz,$$

chaque pôle σ contribuant ainsi pour $(Res_\sigma f)\dfrac{n^\sigma}{\Gamma(\sigma)}$ lorsqu'il est simple, $(Res_\sigma f)\dfrac{n^\sigma \log n}{\Gamma(\sigma)}$ lorsqu'il est double, etc.
Il faut donc extrapoler la fonction $f(k)$ au plan complexe, ce qui se fait par des méthodes à la Weierstrass. Dans le cas de la longueur de cheminement, on définira ainsi $Q(u) = \prod_{k>0}(1-u/2^k)$, de telle sorte que $f(z) = Q(1)/Q(2^{-z})$, dont les pôles (à une translation près) sont en $k \pm 2il\pi$, $k \leq 0$, ce qui conduit aux formules du théorème 3.

3 Les arbres digitaux paginés

Dans ce cas traité par Flajolet et Richmond, on place dans chaque nœud jusqu'à b clefs. Le cas $b = 0$ correspond donc aux *tries* et le cas $b = 1$ aux arbres digitaux. La méthode de résolution précédemment utilisée conduit à des équations de type III sur la s.g.e, qui conduisent par passage aux séries génératrices de Poisson à des récurrences malheureusement non linéaires dès que $b > 1$. Le résultat principal donne un comportement du type déjà observé pour le nombre de sommets de la structure.

Théorème 4 [Flajolet, Richmond] : *Le nombre moyen de sommets dans les arbres b-paginés vaut*

$$s_n = n(q_0 + R(\log_2 n)) + O(\sqrt{n}),$$

où $q_0 = \dfrac{1}{\log 2}\displaystyle\int_0^\infty \frac{(1+t)^{b-1}dt}{((1+t)(1+t/2)(1+t/4)\ldots)^b}$, *et* $R(u)$ *est périodique et développable en série de Fourier.*

Pour $b = 2$, q_0 est une q-fonction de Bessel ; quand b tend vers l'infini, $q_0 \approx 1/(b \log 2)$; et il est bien sûr toujours possible d'estimer numériquement q_0 : on observe que le taux moyen de remplissage est de l'ordre de 70 %.

La méthode est aussi fondée sur le passage aux séries de Poisson, mais cette fois sur les séries génératrices *ordinaires* ! On a alors, en notant $F(z)$ et $G(z)$ les s.g.o. de f_n et g_n, les correspondances suivantes :

$$f_n = \sum_{k=0}^{n} \binom{n}{k} g_k \iff F(z) = \frac{1}{1-z} G(\frac{z}{1-z}),$$

et (bien sûr)

$$f_n = g_{n-b} \iff F(z) = z^b G(z).$$

La récurrence des f_n se transpose heureusement sur les g_n comme le montre l'énoncé suivant.

Proposition : *La s.g.o.* $F(z)$ *du nombre moyen de sommets dans les arbres b-paginés vaut*

$$F(z) = \frac{1}{1-z} G(\frac{z}{1-z}),$$

où $G(t)$ *est donné par*

$$G(t).(1+t)^b = P(t) + 2t^b G(t/2)$$

avec $P(t) = t(1+t)^{b-1}$.

On dispose alors d'une forme explicite pour $G(t)$:

$$G(t) = \frac{P(t)}{(1+t)^b} + \frac{2t^b}{1+t^b}\frac{P(t/2)}{(1+t/2)^b} + \frac{2^2 t^b (t/2)^b}{(1+t)(1+t/2)^b}\frac{P(t/4)}{(1+t/4)^b} + \ldots,$$

ou encore, pour retrouver une forme déjà utilisée,

$$G(t) = \sum_{j\geq 0} 2^j P(2^j t)(\frac{Q(t/2)}{Q(2^j t)})^b, \qquad \text{avec} Q(u) = (1+u)(1+u/2)(1+u/4)...$$

Il reste alors à faire l'analyse du comportement de $G(t)$ pour $t = z/(1-z)$ tendant vers l'infini. Cette analyse se fait de nouveau par transformation de Mellin et conduit aux expressions du théorème 4.

References

[1] N. G. De Bruijn, D. E. Knuth, and S. O. Rice. The average height of planted plane trees. In R. C. Read, editor, *Graph Theory and Computing*, pages 15–22. Academic Press, 1972.

[2] P. Flajolet and R. Sedgewick. Digital search trees revisited. *SIAM Journal on Computing*, 15(3):748–767, August 1986.

[3] Ph. Flajolet and B. Richmond. Generalized digital trees and their difference–differential equations. *Random Structures and Algorithms*, 3(3):305–320, 1992.

[4] D. E. Knuth. *The Art of Computer Programming*, volume 3 : Sorting and Searching. Addison-Wesley, 1973.

[5] A. G. Konheim and D. J. Newman. A note on growing binary trees. *Discrete Mathematics*, 4:57–63, 1973.