Analyse en moyenne d'algorithmes : tri rapide et arbres de recherche

Résumé : Cette thèse est consacrée à l'analyse des comportements en moyenne et des distributions de probabilité de différentes structures d'arbres de recherche et de l'algorithme de tri rapide "QuickSort" avec ses principales variantes : médiane de k, multipartitionnement, tri partiel. L'objectif est de montrer, dans le cas de structures algorithmiques classiques de l'informatique, l'efficacité de certaines méthodes d'analyse combinatoire et d'analyse asymptotique articulées autour de l'utilisation de séries génératrices de dénombrement.

On caractérise l'équivalence combinatoire entre les deux types de structures étudiées : le tri rapide et la recherche arborescente. On étend en particulier cette équivalence à différentes généralisations telles la sélection avec médiane dans tri (arbres "réduits"), ou le multipartitionnement (arbres s-aires de recherche). On montre alors comment, dans le cadre d'une méthodologie plus large, on peut passer directement d'une description de la structure combinatoire et récurrente du problème à un ensemble d'équations analytiques sur des séries génératrices caractéristiques des coûts d'exécution de l'algorithme.

Pour l'algorithme assez général considéré, on met en place un schéma de résolution qui permet d'obtenir les valeurs moyennes exactes pour différentes fonctions de coût possibles. On intègre de plus dans cette résolution des méthodes asymptotiques issues de l'analyse de singularité (lemmes de transfert). Ces méthodes simplifient l'obtention de résultats pratiques sur les coûts moyens et rendent possible le calcul des moments et la caractérisation des distributions de probabilité de ces coûts.

Pour les distributions des coûts liées à la recherche dans les arbres, on prouve ainsi de façon générale, la convergence vers une loi limite gaussienne. Pour le tri Quicksort, on montre l'existence d'une loi limite pour la distribution du nombre de comparaisons ; on fournit de plus des moyens effectifs de caractériser aussi finement que voulu l'ensemble de la distribution.

La dernière partie de ce travail reprend l'analyse en moyenne du tri dans le cas du modèle avec "clés répétées". Des schémas d'analyse et de résolution légèrement différents donnent des résultats intéressants sur le comportement du tri pour différentes configurations envisagées.

Mots-clés : analyse d'algorithme, coût moyen, quicksort, arbre de recherche, permutation, série génératrice, produit partitionnel, variance, distribution limite, analyse asymptotique.

THÈSE

présentée à

L'ÉCOLE POLYTECHNIQUE

PALAISEAU

pour obtenir le titre de

DOCTEUR DE L'ÉCOLE POLYTECHNIQUE

spécialité INFORMATIQUE

par

Pascal HENNEQUIN

Sujet : Analyse en moyenne d'algorithmes, tri rapide et arbres de recherche

Soutenue le 27 mars 1991, devant le jury composé de :

Patrick COUSOT	Président
Maylis DELEST	
Guy FAYOLLE	
Philippe FLAJOLET	
Jean-Marc STEYAERT	

A Françoise,
à Pascale.

Prélude

Extraits de *Encyclopédie Méthodique, ou par ordre de matières : Mathématiques par MM. d'Alembert, l'abbé Bossut, De la Lande, le Marquis de Condorcet, &c* Paris 1784.

PERMUTATION, f. f. (*Analyse*); on entend, par ce mot, la transposition qu'on fait des parties d'un même tout, pour en tirer les divers arrangements dont elles sont suseptibles entre elles. Comme si l'on cherchait en combien de façons différentes on peut disposer les lettres d'un mot, les chiffres qui expriment un nombre, les personnes qui composent une assemblée, &c. Il ne faut pas confondre la *permutation* avec la combinaison. Dans celle-ci, le tout est en quelque sorte démembré, & l'on en prend les différentes parties 1 à 1, 2 à 2, &c. Dans celle-là, le tout conserve toujours son intégrité, & l'on ne fait que faire changer d'ordre aux différents parties qui le constituent.

Pour trouver toutes les permutations possibles d'un nombre quelconque de termes, il ne s'agit que d'un procédé très-simple & très-facile, lequel porte en soi sa démonstration. Il est clair qu'un seul terme a ne peut avoir qu'un arrangement. Si l'on ajoute un second b, on peut le mettre devant ou après le a; ce qui donne deux arrangements ba, ab : c'est à-dire 1 (qu'on avait déjà pour le premier cas) $*2$(quantième du nouveau terme). Si l'on prend un 3^{eme} terme c, il peut occuper trois places dans ba, & autant dans ab, ce qui donne deux fois 3 ou six arrangements $cba, bca, bac, cab, acb, abc$; c'est-à-dire 2 (résultat du cas précédent) * 3 (quantième du nouveau terme). Un quatrième terme d pourra occuper quatre places dans chacun de ces six derniers arrangements; ce qui en donnera 4 fois six ou 24 nouveaux, c'est-à-dire 6 (résultat du cas précédent) * 4 (quantième du nouveau terme). On voit, sans qu'il soit besoin de pousser plus loin l'induction, qu'un cinquième terme e donnerait $24 * 5$ ou 120 arrangements, & ainsi de suite à l'infini. En général, le nombre des *permutations* pour n termes n'étant que celui de $n-1$ termes fois n, comme celui de $n-1$ termes est celui de $n-2$ termes fois $n-1$, & ainsi de suite en remontant jusqu'a 1; il résulte que, pour trouver de combien de permutations est susceptible un nombre quelconque n de termes, il faut faire le produit continu des termes de la progression naturelle, depuis & y compris 1 jusqu'a ce terme n inclusivement $1 * 2 * 3 * 4 \ldots.. * n$.

On a supposé jusqu'ici qu'aucun des termes dont on cherche les permutations n'était répété, ou, ce qui est la même chose, qu'ils n'avaient tous qu'une seule dimension, & que leur exposant commun était l'unité. Si la chose était autrement, supposons que a représente l'exposant du premier terme, b celui de second, c celui du troisième, & ainsi de suite jusqu'au dernier. D'abord n dans la formule ci-dessus, ne fera plus simplement le nombre de termes, mais la somme de leurs exposants. De plus, cette forme ne doit être considérée que comme le numérateur d'une fraction, à laquelle on donnera pour dénominateur le produit continu d'autant de produits particuliers qu'il y a d'exposants ou de termes; & chacun de ces produits particuliers sera le produit continu des nombres naturels poussé jusqu'à celui inclusivement qui exprime l'exposant du terme correspondant, en sorte que la formule absoluement générale sera,

$$\frac{1 * 2 * 3 * 4 \ldots\ldots. * (a + b + c, \&c.)}{1.2\ldots.a * 1.2\ldots..b * 1.2\ldots..c * \&c.}$$

Quand tous les exposants sont 1, alors leur somme ne difère point du nombre même des termes, & $a * b * c \cdots$ (dans le numérateur)$= n * \cdots$ d'ailleurs, dans le dénominateur, tous les produits particulier étant 1, le produit général est aussi 1, qui peut être négligé; & la seconde formule se change en la première.

Un exemple va donner une idée de l'effet des *permutations.* Il y a 32 cartes dans un jeu de piquet; comme c'est un jeu fort répandu, & qu'on mêle les cartes à chaque coup, il s'est dû, depuis le temps qu'on y joue, former bien des arrangements différents de ces 32 cartes; supposant qu'aucun ne se soit jamais trouvé répété, en sorte que chaque fois qu'on a mêlé les cartes en ait fait un nouveau : on demande si le nombre de tous les arrangements possibles ne devrait pas désormais être épuisé . Bien des gens peut-être ne balanceraient pas à se décider pour l'affirmative; on va voir combien ils se trouveraient loin de leur compte. Supposant tous les individus de l'espèce humaine répandus sur la surface de la terre, sans distinction d'age ni de sexe, devenus joueurs de piquet, & appariés deux à deux, en sorte que chaque couple jouat 400 coups par jour *sous la conduite pesée :* il faudra à tous ces joueurs réunis plus de 18 mille milliards de millions de siécles, pour épuiser tous les changements d'ordre possibles des 32 cartes & la démonstration en est facile; 400 coups par jour, en sont par an 146000, par siécle 146000.000, par millions de siécles 14600000.000000. D'un autre coté, supposant deux milliards ou deux mille millions d'hommes sur la terre, ce sera 1.000.000.000 couples de joueurs qu'ils faut multiplier par le dernier nombre ci-dessus, on aura 14.600.000.000.000.000.000.000(A) Maintenant le nombre des permutations compètent à 32 termes, se trouve 263.130.836.933.693.530.167.218.012.160.000.000(B) Si donc on divise le nombre B par le nombre A, le quotient indiquera combien de millions de siécles il faudrait à tous ces joueurs, pour parvenir au but proposé. Or le nombre B ayant 36 chiffres, tandisque le nombre A n'en a que 23, dont le premier plus petit que le premier du nombre B ; le quotient en aura $36 - 23 * 1$, ou 13, dont les deux premiers seront 18. Ce quotient excédera donc 18 mille milliards; & il ne faut pas d'ailleurs perdre de vue que les unités aux-quelles se rapportent ces 18 mille milliards, sont non des années, mais des millions de siécles.

Dans le temps que les anagrammes étaient en honneur & faisaient partie du bel esprit, on voit que, sans nul génie, mais avec beaucoup de loisir & autant de patience, il était aisé de se faire, à cet égard, une réputation; en effet, en suivant avec quelqu'attention le procédé expliqué plus haut, on était assuré de trouver par ordre tous les arrangements possibles des lettres d'un ou de plusieurs mots, tans qu'il en pût échapper un seul, après quoi il ne resterait plus qu'à choisir ceux qui formaient un sens convenable au but qu'on se proposait.

Mais l'usage des permutations ne se borne pas aux seules anagrammes, elles partagent avec les combinaisons l'honneur de la solution de plusieurs problêmes curieux, de ceux en particulier où il s'agit d'estimer les hasards. *V. Combinaison, Alternation, &c.* Article de M. Rallier des Ourmes.

ALGORITHME, f. m. *terme Arabe*, employé par quelques auteurs, & singulièrement par les espagnols, pour signifier *la pratique de l'Algèbre.* Il se prend aussi quelquefois pour l'*Arithmétique par chiffres.* L'*algorithme*, selon la force du mot, signifie proprement l'*art de supputer avec justesse & facilité :* il comprend les règles de l'Arithmétique vulgaire. Ainsi, l'on dit l'*algorithme* des entiers, l'*algorithme* des fractions, l'*algorithme* des nombres sourds, &c. Le même mot se prend, en général, pour désigner la méthode & la notation de toute espèce de calcul. En ce sens, on dit l'*algorithme du calcul intégral*, l'*algorithme du calcul exponentiel*, l'*algorithme du calcul des sinus*, &c.