27

Analyse des arbres suffixes par motif coulissant

Philippe Jacquet
INRIA, Rocquencourt

[résumé par Danièle Gardy]

Dans cet exposé, Ph. Jacquet présente un travail effectué en commun avec W. Szpankowski, consacré à un modèle de *trie* (arbre digital) où les clés, à la différence du modèle usuel, sont fortement corrélées [2, 1].

1 Rappels et définitions

1.1 Tries

Les *tries* sont habituellement construits sur des clés qui sont des suites infinies de bits 0 et 1, ou plus généralement des mots infinis sur un alphabet fini A. Notons une propriété intéressante pour la suite: *soit σ le chemin conduisant de la racine à la feuille où est stockée la donnée X_i; σ est le plus petit préfixe de X_i qui ne soit préfixe d'aucune autre clé X_j.*
La plupart des études sur les tries ont été faites en supposant que les clés qu'ils contiennent sont indépendantes (voir par exemple les résultats rassemblés dans [4, p. 488]). Cependant, certains problèmes sur les mots conduisent à s'intéresser à des tries construits sur des clés fortement corrélées, obtenues par exemple en prenant les suffixes d'un même mot. On parle alors d'*arbre suffixe*. Ce type d'arbre apparaît aussi dans des problèmes liés à la compression de données, ainsi l'algorithme de Lempel et Ziv, qui est à la base de la commande *compress* en Unix.

1.2 Arbres suffixes

Soit un alphabet A, et soit ω un mot de A^*; les *suffixes* de ω sont obtenus en supprimant successivement les premières lettres de ω. Par exemple, prenons $\omega = 01100111...$; les quatre premiers suffixes de ω sont ω lui-même, $1100111...$, $100111...$ et $00111...$ Un *arbre suffixe* de taille n est défini par rapport à un mot ω; c'est le trie formé sur les n premiers suffixes de ω.

1.3 Modèle probabiliste

Les clés sont construites sur un alphabet A de taille V, dont les lettres n'ont pas nécessairement toutes la même probabilité. Un mot ω est un élément de A^*, dont les lettres sont indépendantes (modèle de Bernoulli). Sur ce mot ω, on construit alors un arbre suffixe en prenant les n premiers suffixes.

2 Résultats

Il y a une abondante littérature sur les tries; on pourra se reporter entre autres à l'ouvrage [3] pour une synthèse récente. Il est donc naturel de chercher à s'y rattacher, et à utiliser des résultats déjà connus. L'approche de Ph. Jacquet et W. Szpankowski consiste justement à montrer la proximité de leur modèle avec le modèle classique d'un trie construit sur des clés indépendantes, en termes de fonctions génératrices. Leur résultat fondamental est le suivant:

$$E_{\text{suffixe}}[u^{|\text{chemin}|}] = E_{\text{trie}}[u^{|\text{chemin}|}] + O(n^{-\epsilon}).$$

Ils en déduisent la distribution de la hauteur d'une feuille de l'arbre, i.e. la longueur d'un suffixe d'un mot aléatoire ω: soit D_n la longueur moyenne d'un chemin vers un suffixe quelconque dans l'arbre suffixe d'ordre n construit sur un mot aléatoire ω, et posons $h_1 = -\sum_i p_i \log p_i$ et $h_2 = \sum_i p_i (\log p_i)^2$. Alors

$$E[D_n] = \frac{1}{h_1}(\log n + \gamma + \frac{h_2}{2h_1}) + P_1(\log n) + O(n^{-\epsilon}).$$

De plus, il est possible d'obtenir la variance et la loi limite de D_n, qui diffèrent suivant que les lettres de l'alphabet A sont équiprobables ou non:

- Dans le cas où les lettres de A n'ont pas toutes la même probabilité, la variance est d'ordre $\log n$:
$$var(D_n) = \frac{h_2 - h_1^2}{h_1^3} \log n + C + P_2(\log n) + O(n^{-\epsilon}).$$
De plus, la distribution de D_n est asymptotiquement gaussienne.

- Dans le cas symétrique, $p_i = 1/V$ pour tout i, donc $h_2 = h_1^2$, et la variance tend alors vers une limite finie non nulle:
$$var(D_n) = \frac{\pi^2}{6(\log V)^2} + \frac{1}{12} + P_3(\log n) + O(n^{-\epsilon}).$$
Il existe là aussi une loi limite pour D_n, qui est ici de type double exponentielle:
$$\text{Proba}\{D_n \leq \log_V n + x\} \to exp(-V^{-x}).$$

3 Preuve du théorème fondamental

Soit $\mathcal{C}$ un ensemble de n clés (corrélées ou non): $\mathcal{C} = \{X_1, X_2, \ldots, X_n\}$.

Définition 1 *Soit σ un mot fini; $\langle\sigma\rangle_{\mathcal{C}}$ est l'ensemble des indices i tels que σ soit un préfixe de la clé X_i de $\mathcal{C}$.*

Pour un trie ou un arbre suffixe construit sur les clés d'un ensemble $\mathcal{C}$, notons $D_n^{\mathcal{C}}[i]$ la longueur du chemin allant de la racine vers la feuille contenant la clé X_i. On a la propriété suivante:

$D_n^{\mathcal{C}}[i] \leq k$ si et seulement s'il existe une chaîne σ de longueur k qui n'est préfixe que de X_i, i.e. telle que $\langle\sigma\rangle_{\mathcal{C}} = \{i\}$.

Les événements Proba($\langle\sigma\rangle_{\mathcal{C}} = \{i\}$) étant disjoints, on a:

$$\text{Proba}(D_n^{\mathcal{C}}[i] \le k) = \sum_{|\sigma|=k} \text{Proba}(\langle\sigma\rangle_{\mathcal{C}} = \{i\}).$$

Soit $D_n^{\mathcal{C}}$ la longueur moyenne du chemin de la racine jusqu'à une clé aléatoire de l'ensemble $\mathcal{C}$: $D_n^{\mathcal{C}}$ prend la valeur $D_n^{\mathcal{C}}[i]$ avec une probabilité $1/n$, quel que soit i. On montre (en conditionnant par rapport à l'indice i retenu) que

$$\text{Proba}(D_n^{\mathcal{C}} \le k) = (1/n) \sum_{i=1}^{n} \text{Proba}(D_n^{\mathcal{C}}[i] \le k).$$

Posons $D_n(u) = E[u^{D_n}]$; on a $D_n(u) = \sum_{\mathcal{C}} \text{Proba}(\mathcal{C}) E[u^{D_n^{\mathcal{C}}}]$. En utilisant l'abréviation

$$f(\sigma, i) = \sum_{\mathcal{C}} \text{Proba}(\mathcal{C})\ \text{Proba}(\langle\sigma\rangle_{\mathcal{C}} = \{i\}),$$

où la somme est prise sur tous les ensembles de clés $\mathcal{C}$ admissibles pour le problème étudié, on trouve que

$$D_n(u) = \frac{1}{n}(1-u) \sum_{\sigma} u^{|\sigma|} \sum_{i=1}^{n} f(\sigma, i),$$

et l'étape suivante est de calculer $\sum_i f(\sigma, i)$. Soit $\mathcal{L}$ un sous-ensemble de $\{1, 2, \ldots, n\}$, et soit

$$P(\mathcal{L}, \sigma) = \sum_{\mathcal{C}} \text{Proba}(\mathcal{C})\ \text{Proba}(\mathcal{L} \subset \langle\sigma\rangle_{\mathcal{C}}).$$

Par un argument d'inclusion-exclusion, on montre que, pour tout ensemble de clés $\mathcal{C}$,

$$\text{Proba}(\langle\sigma\rangle_{\mathcal{C}} = \{i\}) = \sum_{j=1}^{n} (-1)^{j+1} \sum_{|\mathcal{L}|=j,\ i\in\mathcal{L}} \text{Proba}(\mathcal{L} \subset \langle\sigma\rangle_{\mathcal{C}}).$$

En sommant sur les ensembles $\mathcal{C}$, on obtient

$$f(\sigma, i) = \sum_{j=1}^{n} (-1)^{j+1} \sum_{|\mathcal{L}|=j,\ i\in\mathcal{L}} P(\mathcal{L}, \sigma).$$

Pour obtenir $D_n(u)$, on peut alors calculer $P(\mathcal{L}, \sigma)$, puis $f(\sigma, i)$, ce qui va être fait ci-dessous d'abord pour des tries construits sur des clés indépendantes, puis pour des arbres suffixes.

3.1 Tries sur des clés indépendantes

Supposons que $\sigma = a_1 a_2 \cdots a_k$, et notons $p(\sigma) = p_{a_1} p_{a_2} \ldots p_{a_k}$. On a alors $P(\mathcal{L}, \sigma) = p(\sigma)^{|\mathcal{L}|}$ et donc $f(\sigma, i) = p(\sigma)(1 - p(\sigma))^{n-1}$. On en tire:

$$D_n(u) = \frac{1}{n}(1-u) \sum_{\sigma} n\ p(\sigma)(1 - p(\sigma))^{n-1} u^{|\sigma|}.$$

Définissons $D_T(z, u) = \sum_n n D_n(u) z^n$:

$$D_T(z, u) = (1-u) \sum_{\sigma} u^{|\sigma|} \frac{p(\sigma) z}{(1 - z + p(\sigma) z)^2}. \tag{1}$$

3.2 Arbres suffixes

Par un raisonnement analogue, bien que plus complexe, en posant $D_S(z,u) = \sum_n n D_n(u) z^n$, où cette fois $D_n(u) = E(u^{D_n})$ est calculé sur les arbres suffixes, on obtient:

$$D_S(z,u) = (1-u)\sum_\sigma \frac{p(\sigma)z}{\left((1-z)(1+a_\sigma(z)) + p(\sigma)z^{|\sigma|}\right)^2} u^{|\sigma|}, \tag{2}$$

où a_σ est le polynôme d'auto-corrélation de σ, défini par la somme suivante, avec σ' décrivant l'ensemble des préfixes de σ qui sont aussi suffixes de σ (et bien sûr $\sigma' \neq \sigma$):

$$a_\sigma(z) = \sum_{\sigma'} \frac{p(\sigma)}{p(\sigma')} z^{|\sigma|-|\sigma'|}.$$

Par exemple [1, p. 16], soit $\sigma = abaaba$ de longueur 6. Alors l'ensemble des mots qui sont à la fois préfixes et suffixes de σ est $\mathcal{E} = \{a, aba\}$, et

$$a_\sigma(z) = \sum_{\sigma' \in \mathcal{E}} \frac{p(\sigma)}{p(\sigma')} z^{|\sigma|-|\sigma'|} = p_a^2 p_b z^3 + p_a^3 p_b^2 z^5.$$

L'idée de la démonstration conduisant à la formule (2) est la suivante: lorsque les clés sont des suffixes d'un même texte (supposé infini), et lorsque ces suffixes débutent à des positions distantes d'au moins k, k étant la longueur de σ, alors ces suffixes peuvent être considérés comme indépendants. On regroupe alors les premières positions, jusqu'à ce qu'on ait un trou de k; les suffixes suivants sont traités indépendamment des précédents.

3.3 Comparaison des deux fonctions

A défaut de pouvoir étudier directement la fonction définie par l'équation (2), on va la comparer avec la fonction analogue pour les tries, définie par l'équation (1), et qui a été bien étudiée. Cela conduit à s'intéresser à la différence

$$Q_n(u) = \frac{1}{1-u}(D_n^{\text{trie}}(u) - D_n^{\text{suffixe}}(u)).$$

D'après la formule de Cauchy, $Q_n(u)$ s'exprime par une intégrale:

$$Q_n(u) = \frac{1}{2i\pi n(1-u)} \oint (D_T(z,u) - D_S(z,u)) \frac{dz}{z^{n+1}}.$$

Ici intervient, pour chaque mot fini σ, la racine de plus petit module A_σ du polynôme $R_\sigma(z) = (1-z)(1+a_\sigma(z)) + p(\sigma)z^{|\sigma|}$. Notons $C_\sigma = R'_\sigma(A_\sigma)$ et $D_\sigma = R''_\sigma(A_\sigma)$. On montre que, pour un B "convenable":

$$Q_n(u) = \frac{1}{n}\sum_\sigma u^{|\sigma|} p(\sigma) \left(A_\sigma^{-n}\left(\frac{n}{A_\sigma C_\sigma^2} + \frac{D_\sigma}{C_\sigma^3}\right) - n(1-p(\sigma))^{n-1} \right) + O(B^{-n}),$$

puis que

$$Q_n(u) = g_n(u) + O(\frac{1}{n}),$$

avec

$$g_n(u) = \sum_{\sigma} u^{|\sigma|} p(\sigma) f_\sigma(n)$$

et avec

$$f_\sigma(x) = \frac{A_\sigma^{-x} - e^{-x}}{A_\sigma C_\sigma^2} - \frac{(1-p(\sigma))^x - e^{-x}}{1-p(\sigma)}.$$

Il reste alors à voir qu'il existe $\epsilon > 0$ tel que $g_n(u)$ est $O(n^{-\epsilon})$. Pour cela, on utilise des propriétés qui relient le comportement asymptotique d'une fonction (ici $g_n(u)$, en tant que fonction de n) à la bande fondamentale de sa transformée de Mellin, i.e. à l'ensemble sur lequel cette transformée est "naturellement" définie. On calcule donc, pour chaque σ, la transformée de Mellin de f_σ:

$$f_\sigma^*(s) = \int_0^{+\infty} f_\sigma(x) x^{s-1} ds = \Gamma(s) \left(\frac{(\log A_\sigma)^{-s} - 1}{A_\sigma C_\sigma^2} - \frac{(-\log(1-p(\sigma))^{-s} - 1}{1-p(\sigma)} \right).$$

Pour un motif σ donné, cette transformée est définie sur $]-1, +\infty[$. Lorsqu'on somme sur tous les mots σ, on peut montrer que, pour un certain $\epsilon > 0$, $g^*(s) = \sum_\sigma u^{|\sigma|} p(\sigma) f_\sigma^*(s)$ reste définie sur $]-1, \epsilon[$. D'où la majoration cherchée: $g_n(u) = O(n^{-\epsilon})$, à quelques détails techniques près.

Références

[1] Ph. Jacquet and W. Szpankowski. Autocorrelation on words and its application: Analysis of suffix trees by string-ruler approach. Technical Report 1106, Institut National de Recherche en Informatique et en Automatique, October 1991. Accepted for publication in Journal of Combinatorial Theory, Series A.

[2] Ph. Jacquet and W. Szpankowski. What can we learn about suffix trees from independent tries? In *Second Workshop on Algorithms and Data Structures*, Lecture Notes in Computer Science, pages 228–239, Ottawa (Canada), August 1991. Springer Verlag.

[3] H. M. Mahmoud. *Evolution of Random Search Trees*. John Wiley, 1992.

[4] J. Vitter and Ph. Flajolet. Analysis of Algorithms and Data Structures. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume A: Algorithms and Complexity, chapter 9, pages 431–524. North Holland, 1990.