УДК 519.212.2

Число пар одинаково помеченных вхождений заданного поддерева в q-ичное дерево со случайными метками вершин

А. М. Зубков[*], **В. И. Круглов**[*]

[*] *Математический институт им. В.А. Стеклова Российской академии наук ул. Губкина, д.8, Москва, Россия, 119991*

Аннотация. Рассматривается полное q-ичное дерево, каждой вершине которого случайно, равновероятно и независимо от остальных вершин присвоена метка из конечного алфавита. Получена асимптотическая формула для среднего числа пар одинаково помеченных вхождений заданного поддерева, сформулирована теорема о сходимости распределения числа таких пар к пуассоновскому распределению.

Ключевые слова: q-ичные деревья с помеченными вершинами, суммы зависимых индикаторов, пуассоновская аппроксимация.

1. Основная часть

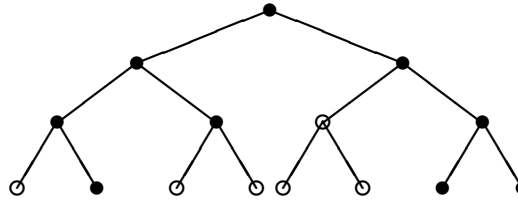
Пусть T_q^n — полное q-ичное дерево высоты n. Корень дерева будем обозначать символом $*$ и считать, что он образует слой вершин $I^{(0)}$. Для каждого $k = 1, \dots, n$ слой $I^{(k)}$ состоит из q^k вершин $\mathbf{i} = [i_1, i_2, \dots, i_k]$, где $i_1, i_2, \dots, i_k \in \{0, 1, \dots, q-1\}$. Из корня $*$ выходит q ребер в вершины $[0], [1], \dots, [q-1] \in I^{(1)}$, при $k = 1, \dots, n-1$ из каждой вершины $\mathbf{i} = [i_1, i_2, \dots, i_k]$ слоя $I^{(k)}$ выходит q ребер, соединяющих ее с вершинами $\mathbf{i}_j^+ = [\mathbf{i}, j] = [i_1, i_2, \dots, i_k, j]$, $j = 0, 1, \dots, q-1$, слоя $I^{(k+1)}$. В каждую вершину $\mathbf{i} = [i_1, i_2, \dots, i_k]$ при $k > 1$ входит ребро из вершины $\mathbf{i}^- = [i_1, i_2, \dots, i_{k-1}]$, а при $k = 1$ — из корня $* = [0]^- = [1]^- = \dots = [q-1]^-$. Для вершины $\mathbf{i} \in I^{(k)}$ будем считать, что ее «высота» $h(\mathbf{i}) = k$. Если $\mathbf{i} = [i_1, i_2, \dots, i_k]$, $\mathbf{j} = [j_1, j_2, \dots, j_m]$, то $[\mathbf{i}, \mathbf{j}] \stackrel{\text{def}}{=} [i_1, i_2, \dots, i_k, j_1, j_2, \dots, j_m]$, $[\mathbf{i}, *] \stackrel{\text{def}}{=} [i_1, i_2, \dots, i_k]$.

На множестве вершин дерева T_q^n введем лексикографический порядок, считая, что $\mathbf{i} \prec \mathbf{j}$, если либо $h(\mathbf{i}) < h(\mathbf{j})$, либо $h(\mathbf{i}) = h(\mathbf{j}) = k \in \{1, 2, \dots, n\}$, $\mathbf{i} = (i_1, \dots, i_k)$, $\mathbf{j} = (j_1, \dots, j_k)$ и $\sum_{m=1}^k i_m q^{k-m} < \sum_{m=1}^k j_m q^{k-m}$.

Определение 1. *Шаблоном* будем называть поддерево B дерева T_q^n с корнем $*$ и $|B|$ вершинами $\mathbf{i}_0^{[B]} = * \prec \mathbf{i}_1^{[B]} \prec \dots \prec \mathbf{i}_{|B|-1}^{[B]}$. Если $\mathbf{i}_{|B|-1}^{[B]} \in I^{(h)}$, то будем называть h *высотой* $h(B)$ шаблона B. При $\mathbf{j} \in I^{(k)}$, $0 \leqslant k \leqslant n - h(B)$, *вложением* $B(\mathbf{j})$ шаблона B в дерево

T_q^n называется поддерево дерева T_q^n, имеющее вершины $\mathbf{j} = [\mathbf{j}, \mathbf{i}_0^{[B]}] \prec [\mathbf{j}, \mathbf{i}_1^{[B]}] \prec [\mathbf{j}, \mathbf{i}_2^{[B]}] \prec \ldots \prec [\mathbf{j}, \mathbf{i}_{|B|-1}^{[B]}]$.

Пример: на рисунке для случая $q = 2$ изображен шаблон B с $h(B) = 3$, $|B| = 9$ и вершинами $\mathbf{i}_0^{[B]} = * \prec \mathbf{i}_1^{[B]} = [0] \prec \mathbf{i}_2^{[B]} = [1] \prec \mathbf{i}_3^{[B]} = [0,0] \prec \mathbf{i}_4^{[B]} = [0,1] \prec \mathbf{i}_5^{[B]} = [1,1] \prec \mathbf{i}_6^{[B]} = [0,0,1] \prec \mathbf{i}_7^{[B]} = [1,1,0] \prec \mathbf{i}_8^{[B]} = [1,1,1]$.



Пусть каждой вершине $\mathbf{i}$ дерева T_q^n присвоена случайная метка $m(\mathbf{i})$, принимающая значения в множестве $\{1, \ldots, d\}$, причем значения $m(\mathbf{i})$, $\mathbf{i} \in T_q^n$, независимы в совокупности и $\mathbf{P}\{m(\mathbf{i}) = j\} = \frac{1}{d}, j \in \{1, \ldots, d\}$, для всех $\mathbf{i} \in T_q^n$. Каждому вложению $B(\mathbf{j})$ шаблона B соответствует упорядоченный набор

$$M(B(\mathbf{j})) = \left(m([\mathbf{j}, \mathbf{i}_k^{[B]}]), k = 0, 1, \ldots, |B| - 1\right)$$

случайных меток вершин поддерева $B(\mathbf{j})$.

Очевидно, что если для некоторых вершин $\mathbf{g_1}, \ldots, \mathbf{g_s}$ поддеревья $B(\mathbf{g_1}), \ldots, B(\mathbf{g_s})$ попарно не пересекаются, то соответствующие им совокупности меток $M(B(\mathbf{g_1}), \ldots, M(B(\mathbf{g_s}))$ взаимно независимы и имеют равновероятное распределение на множестве $\{1, \ldots, d\}^{|B|}$.

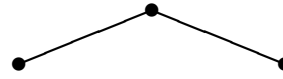
Задачи о появлении или непоявлении заданных комбинаций знаков в строках рассматривались, например, в [7], задачи о появлении заданных поддеревьев в деревьях — в [8, 9]. Задачи о появлении одинаковых комбинаций в случайных дискретных последовательностях рассматривались, например, в [1, 3–5, 10]. Первые результаты о совпадениях меток на цепочках вершин в двоичном дереве были получены авторами настоящей работы в [2]. Такие характеристики имеют естественные интерпретации, например, для генеалогических деревьев и для повторяющихся вычислений при переборе с ветвлением вариантов. Здесь изучаются распределения чисел непродолжаемых к корню совпадений наборов меток в разных вложениях заданного шаблона B

(зависящего от n), т. е. сумм индикаторов вида

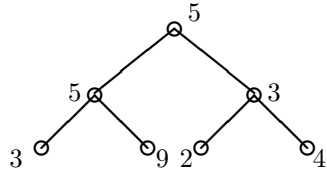
$$X_{\mathbf{i},\mathbf{j},B} = \mathbb{I}\{M(B(\mathbf{i}) = M(B(\mathbf{j}), m(\mathbf{i}^-) \neq m(\mathbf{j}^-)\}, \quad \mathbf{i},\mathbf{j} \in \bigsqcup_{k=0}^{n-l+1} I^{(k)},\ \mathbf{i} \prec \mathbf{j}$$

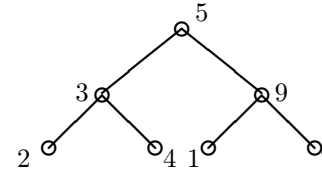
(при $\mathbf{i} = *$ условие $m(\mathbf{i}^-) \neq m(\mathbf{j}^-)$ считается выполненным автоматически).

Пример: для двоичного дерева рассмотрим шаблон B, для которого $h(B) = 2$ и $|B| = 3$,



и два фрагмента двоичного дерева со следующими метками вершин:





причем мы предполагаем, что вершина, предшествующая корневой вершине правого поддерева, имеет метку, отличную от 5. Тогда в рассмотренных фрагментах имеет место одно непродолжаемое к корню совпадение меток, рассматриваемых в соответствии с шаблоном B, которое образовано вершинами с метками 5, 3 и 9, а также еще одно совпадение, образованное вершинами с метками 3, 2 и 4, которое может быть продолжено к корню и поэтому не учитывается.

Для каждого натурального n и шаблона $B = B_n$ высоты $h_n = h(B_n)$ мы рассматриваем случайные величины

$$V_{n,B_n} = \sum_{\mathbf{i},\mathbf{j} \in T_q^{n-h_n}\colon \mathbf{i} \prec \mathbf{j}} X_{\mathbf{i},\mathbf{j},B_n} \quad \text{и} \quad \widetilde{V}_{n,B_n} = \sum_{(\mathbf{i},\mathbf{j}) \in \mathcal{P}_{n,B_n}} X_{\mathbf{i},\mathbf{j},B_n},$$

где

$$\mathcal{P}_{n,B_n} = \left\{(\mathbf{i},\mathbf{j})\colon \mathbf{i},\mathbf{j} \in T_q^{n-h_n}, * \prec \mathbf{i} \prec \mathbf{j},\ \mathbf{j} \neq [\mathbf{i},\mathbf{z}]\, \forall \mathbf{z}\colon h(\mathbf{z}) \leqslant h_n + 1\right\}.$$

Ограничение одним шаблоном B_n, условие $m(\mathbf{i}^-) \neq m(\mathbf{j}^-)$ и выбор множества $\mathcal{P}_{n,B_n}$ сделаны для того, чтобы упростить формулировку

и доказательство теоремы о сходимости распределений сумм V_{n,B_n} и $\widetilde{V}_{n,B_n}$ к распределению Пуассона при $n \to \infty$.

Лексикографический порядок на множестве вершин дерева T_q^n обладает, в частности, следующим свойством: если $\mathbf{i} \prec \mathbf{j}$, то такое же отношение порядка существует между любыми двумя соответствующими вершинами вложений $B(\mathbf{i})$ и $B(\mathbf{j})$ шаблона B: $[\mathbf{i}, \mathbf{i}_k^{[B]}] \prec [\mathbf{j}, \mathbf{i}_k^{[B]}]$, $1 \leqslant k < |B|$. Поэтому в последовательности $m(\mathbf{i}^-) \neq m(\mathbf{j}^-)$, $m([\mathbf{i}, \mathbf{i}_k^{[B]}]) = m([\mathbf{j}, \mathbf{i}_k^{[B]}]), k = 0, 1, \ldots, |B| - 1$, каждое соотношение содержит хотя бы одну новую вершину поддерева $B(\mathbf{j})$. Из этого свойства и из предположения о независимости и равновероятности меток следует, что условная вероятность выполнения каждого соотношения при условии, что выполнены все предыдущие, равна безусловной вероятности выполнения этого соотношения, т. е. что при $\mathbf{i} \prec \mathbf{j}$

$$\mathbf{E}X_{\mathbf{i},\mathbf{j},B} = \mathbf{E}\mathbb{I}\{m([\mathbf{i}, \mathbf{i}_k^{[B]}]) = m([\mathbf{j}, \mathbf{i}_k^{[B]}]), 0 \leqslant k < |B|\}\mathbb{I}\{m(\mathbf{i}^-) \neq m(\mathbf{j}^-)\} =$$

$$= \mathbf{P}\{m(\mathbf{i}^-) \neq m(\mathbf{j}^-)\} \prod_{k=0}^{|B|-1} \mathbf{P}\{m([\mathbf{i}, \mathbf{i}_k^{[B]}]) = m([\mathbf{j}, \mathbf{i}_k^{[B]}])\} =$$

$$= \begin{cases} \dfrac{d-1}{d^{|B|+1}}, & \text{если } \mathbf{i} \neq *, \\ \dfrac{1}{d^{|B|}}, & \text{если } \mathbf{i} = *. \end{cases}$$

Теорема 1. *Если $n, h_n = h(B_n) \to \infty$ так, что $n - h_n \to \infty$, то*

$$\mathbf{E}V_{n,B_n} = \frac{d-1}{d^{|B_n|+1}} \frac{q^{2(n-h_n+1)}}{2(q-1)^2} (1 + o(1)),$$

$$0 \leqslant \mathbf{E}V_{n,B_n} - \mathbf{E}\widetilde{V}_{n,B_n} = o(\mathbf{E}V_{n,B_n}).$$

Следствие 1. *Если $n, h_n = h(B_n) \to \infty$ так, что $n - h_n \to \infty$ и величина $\mathbf{E}V_{n,B_n}$ остается ограниченной, то $\mathbf{P}\{\widetilde{V}_{n,B_n} = V_{n,B_n}\} \to 1$.*

Применяя метод Чена-Стейна (см., например, [6]), можно получить достаточное условие для стремления к нулю расстояния по вариации между распределением $\mathcal{L}(\widetilde{V}_{n,B_n})$ случайной величины $\widetilde{V}_{n,B_n}$ и аппроксимирующим пуассоновским распределением $Pois(\mathbf{E}\widetilde{V}_{n,B_n})$:

$$d_{\text{tv}}(\mathcal{L}(\widetilde{V}_{n,B_n}), Pois(\mathbf{E}\widetilde{V}_{n,B_n})) =$$

$$= \frac{1}{2} \sum_{k=0}^{\infty} \left| \mathbf{P}\{\widetilde{V}_{n,B_n} = k\} - \mathbf{P}\{Pois(\mathbf{E}\widetilde{V}_{n,B_n}) = k\} \right|.$$

Теорема 2. *Если* $n, h_n = h(B_n) \to \infty$ *так, что* $n - 2h_n \to \infty$, *то для некоторой функции* $\varepsilon(n) = o(1), n \to \infty$, *справедливы неравенства*

$$d_{\mathrm{tv}}(\mathcal{L}(\widetilde{V}_{n,B_n}), \mathrm{Pois}(\mathbf{E}\widetilde{V}_{n,B_n})) \leq \frac{16\left(1 - e^{-\mathbf{E}\widetilde{V}_{n,B_n}}\right)\mathbf{E}\widetilde{V}_{n,B_n}}{q^{n-2h_n-1}}\left(1 + \varepsilon(n)\right);$$

если дополнительно $q^n = o(d^{|B_n|})$, *то*

$$d_{\mathrm{tv}}(\mathcal{L}(\widetilde{V}_{n,B_n}), \mathrm{Pois}(\mathbf{E}\widetilde{V}_{n,B_n})) \to 0.$$

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 14-50-00005).

Литература

1. *Зубков А. М., Михайлов В. Г.* Предельные распределения случайных величин, связанных с длинными повторениями в последовательности независимых испытаний // Теория вероятн. и ее примен. — 1974. — Т. 19, вып. 1. — С. 173–181.
2. *Зубков А. М., Круглов В. И.* Повторения цепочек на бинарном дереве со случайными метками вершин // Дискрет. матем. — 2015. — Т. 27, вып. 4. — С. 38—48.
3. *Михайлов В. Г.* Оценка точности сложной пуассоновской аппроксимации для распределения числа совпадающих цепочек // Теория вероятн. и ее примен. — 2001. — Т. 46, вып. 4. — С. 713–723.
4. *Михайлов В. Г.* Оценки точности пуассоновской аппроксимации для распределения числа серий повторений длинных цепочек в цепи Маркова // Дискрет. матем. — 2015. — Т. 27, вып. 4. — С. 67–78.
5. *Михайлов В. Г.* О вероятности наличия в случайной последовательности цепочек с одинаковой структурой // Дискрет. матем. — 2016. — Т. 28, вып. 3. — С. 97–110.
6. *Erhardsson T.* Stein's method for Poisson and compound Poisson approximation // Barbour A. D., Chen L. H. Y. (ed.) "An introduction to Stein's method". — Singapore Univ. Press, 2005. — P. 61–113.
7. *Guibas L. J., Odlyzko A. M.* Long repetitive patterns in random sequences // Z. Wahrscheinlichkeitstheorie verw. Geb. — 1980. — Vol. 53. — P. 241–262.
8. *Hoffmann C. M., O'Donnell M. J.* Pattern matching in trees // J. ACM. — 1982. — Vol. 29, no. 1. — P. 68-95.
9. *Steyaert J.-M., Flajolet P.* Patterns and pattern-matching in trees: an analysis // Inf. & Control. — 1983. — Vol. 58, no. 1. — P. 19–58.

10. *Karlin S., Ost F.* Counts of long aligned word matches among random letter sequences // Adv. Appl. Probab. — 1987. — Vol. 19, no. 2. — P. 293–351.

UDC 519.212.2

Number of pairs of identically marked embeddings of given subtree in q-ary tree with randomly marked vertices

A. M. Zubkov[*], **V.I. Kruglov**[*]

[*] *Steklov Mathematical Institute of Russian Academy of Sciences, Gubkina str. 8, Moscow, 119991, Russia*

Let for all vertices of a complete q-ary tree independent random marks are assigned and all marks have uniform distribution on a finite alphabet. We consider pairs of identically marked embeddings of a given subtree. An asymptotic formula for the expectation of the number of such pairs is obtained and the Poisson limit theorem for this number is proposed.

This work is supported by the Russian Science Foundation under grant № 14-50-00005.

Keywords: q-ary trees with marked vertices, sums of dependent indicators, Poisson approximation.