Серия «Математика»
2013. Т. 6, № 3. С. 38—47

Онлайн-доступ к журналу:
http://isu.ru/izvestia

ИЗВЕСТИЯ
Иркутского государственного университета

УДК 519.111

Конечные префиксные коды и финитарные деревья

И. О. Коряков
Уральский федеральный университет

Аннотация. Установлена явная формула для числа всех p-арных k-деревьев с m листьями. Ее анализ приводит к ряду комбинаторных тождеств.

Ключевые слова: Дерево; префиксный код; производящая функция; число Фусса; число Каталана; язык Моцкина.

1. Введение

В настоящей заметке рассматриваются некоторые задачи перечисления указанных в названии объектов. Обе комбинаторные модели (коды и деревья) по существу идентичны (см. ниже лемму 2), чем и объясняется параллелизм их обсуждения. Основной параметр перечисления — мощность кода — оказывается достаточным лишь в случае полных префиксных кодов. Все числа здесь — натуральные; $\langle n\rangle = \{1, 2, \ldots, n\}$, $n \geqslant 0$.

Начнём с азбуки комбинаторной дендрологии. В нашем «дендрарии» *все деревья — корневые* и, значит, определяются отношением «отец–сын». *Отец* — это узел, имеющий сына, *лист* — узел без сыновей. Любое поддерево определяется (и отождествляется с) множеством его узлов; в частности, мощность $|T|$ дерева T — это число узлов, дерево мощности k называется *k-деревом.*

Дерево *упорядочено*, если для каждого узла x множество $S(x)$ его сыновей линейно упорядочено; порядок обычно называют *старшинством* (и рисуют братьев слева направо по убыванию старшинства).

Пусть $p > 0$; *p-арным* назовём дерево, в котором для каждого узла x задана инъекция $\alpha_x : S(x) \to \langle p\rangle$. Такое дерево становится упорядоченным, если для братьев u, v — сыновей узла x — положить $u < v$, когда

$\alpha_x(u) < \alpha_x(v)$. *Финитарными*[1] мы называем p-арные деревья для различных p. На рис. 1 показаны все бинарные (первый ряд) и тернарные 3-деревья (корень — вверху, старший сын — внизу слева, средний — внизу вертикально, младший — внизу справа).

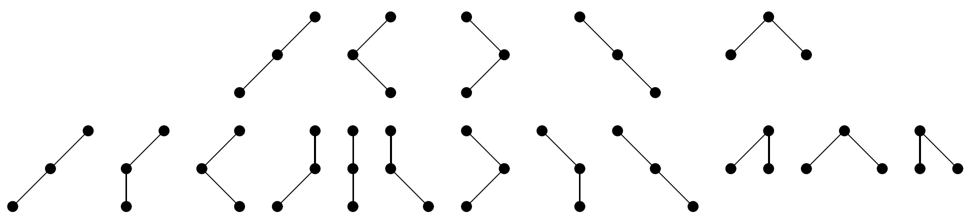

Рис. 1. Все бинарные и тернарные 3-деревья

В индуктивных рассуждениях удобно использовать *рекурсивное представление* p-арного дерева, определяемое следующим образом.

Пусть S — множество сыновей корня, $\alpha : S \to \langle p \rangle$ — инъекция из определения. Для всех $i \in \langle p \rangle$ положим $T_i = \varnothing$, если $\alpha^{-1}(i) = \varnothing$, иначе T_i — наименьшее поддерево, содержащее $s = \alpha^{-1}(i)$ и всех потомков узла s. Тогда мы пишем $T \simeq (T_1, T_2, \ldots, T_p)$.

Полным p-арным деревом назовём непустое p-арное дерево, в котором каждый отец имеет p сыновей. Легко видеть, что p-арное дерево T полно тогда и только тогда, когда либо T состоит из одного корня, либо $T \simeq (T_1, \ldots, T_p)$, где все T_i — полные p-арные деревья.

2. Основные результаты

Обозначим через $\mathcal{T}(p; k)$ множество всех p-арных k-деревьев, а через $\mathcal{CT}(p; k)$ множество всех p-арных деревьев с k отцами. Для $p, k > 0$ положим

$$F(p,k) = \frac{1}{pk+1}\binom{pk+1}{k} = \frac{1}{k}\binom{pk}{k-1};$$

это *число Фусса* (1791 г.). Заметим, что $F(2,k) = C_k = \frac{1}{k+1}\binom{2k}{k}$ — k-е число Каталана. Хорошо известна

Лемма 1. (1) $|\mathcal{T}(p;k)| = |\mathcal{CT}(p;k)| = F(p,k)$;
(2) *число узлов в дереве из* $\mathcal{CT}(p;k)$ *равно* $pk+1$, *а число листьев равно* $(p-1)k+1$.

---
[1] По аналогии с финитарными потоками (веерами) в интуиционистской математике Л. Брауэра.

Доказательство. Равенство $|\mathcal{CT}(p;k)| = F(p,k)$ доказывается либо с помощью производящих функций, либо комбинаторно (см., например, [1], где очевидно и (2)).

«Дефолиация», т. е. удаление всех листьев дерева, является отображением из $\mathcal{CT}(p;k)$ на $\mathcal{T}(p;k)$. Его обратимость легко доказывается индукцией по k с использованием рекурсивного представления. $\square$

Разумеется, перечисление ведётся с точностью до изоморфизма, который определяется очевидным образом.

Пусть A — алфавит и $u, w \in A^*$. Мы пишем $u \leqslant w$, если u — префикс слова w, т. е. $w = uv$ для некоторого $v \in A^*$. Язык $C \subseteq A^*$ — *префиксный код над* A, если $C \cap CA^+ = \varnothing$ (см. [2], гл. 5). Код C назовём p*-арным*, если $|A| = p$. Для каждого $p > 0$ зафиксируем алфавит $A_p = \{a_1, \dots, a_p\}$ и будем рассматривать лишь коды над A_p, считая их изоморфными только при совпадении.

Лемма 2. *Существует биекция $T \mapsto T(C)$ множества всех конечных p-арных префиксных кодов на множество всех конечных p-арных деревьев; при этом слова из C отвечают листьям дерева $T(C)$.*

Доказательство. Пусть C — префиксный код над A_p. Рассмотрим множество $T(C) = \{u \in A_p^* \mid u \leqslant w \in C\}$ префиксов всех кодовых слов и введём на нём структуру p-арного дерева: корнем объявим пустое слово e; если $wa_i \in \in T(C)$, где $a_i \in A_p$, то объявим wa_i сыном узла w; определим инъекцию $\alpha : S(w) \to \langle p \rangle$, полагая $\alpha(wa_i) = i$. Ясно, что C — это множество листьев в $T(C)$. Обратное отображение определяется очевидным образом. $\square$

Из лемм 1 и 2 вытекает

Теорема 3. *Число всех p-арных префиксных кодов C, таких, что $|T(C)| = k$, равно числу Фусса $F(p,k)$.* $\square$

Префиксный код C над алфавитом A называется *полным*, если C максимален в классе всех префиксных кодов над A. Условие полноты можно записать в терминах слов ([2], гл. 5):

$$\forall w \in A^* \, \exists c \in C (w \leqslant c \vee c < w).$$

Лемма 4. *Префиксный код C полон тогда и только, когда финитарное дерево $T(C)$ полно.*

Доказательство. Пусть C — полный префиксный код, но дерево $T(C)$ не полно, т. е. некоторый отец u не имеет, скажем, сына ua, где a — буква. В силу полноты кода $ua \leqslant c$ или $c < ua$ для некоторого $c \in C$. Первое неравенство даёт $ua \in T(C)$, что противоречит предположению; из второго следует $c \leqslant u < c' \in C$, где c' — потомок узла u, что невозможно в силу префиксности C. Поэтому $T(C)$ полно.

Обратно, предположим, что дерево $T(C)$ полно, но код C не полон. Пусть $w = ua$ (где a — буква) — такое слово наименьшей длины, что $w \notin C$ и $C_1 = C \cup \{w\}$ — префиксный код. Тогда $C \cup \{u\}$ не является новым префиксным кодом, т. е. либо $u \in C$, либо $u \notin C$ и тогда множество $C \cup \{u\}$ не обладает свойством префиксности. Первый случай противоречит префиксности C_1. Во втором случае для некоторого $c \in C$ либо $c < u$ (а тогда $c < w$, что опять противоречит префиксности C_1), либо $u < c$. Последнее неравенство означает, что u — отец в $T(C)$ и — в силу полноты $T(C)$ — $w = ua \in T(C)$, что вновь противоречит префиксности C_1. Итак, код C — полный. □

Теорема 5. (1) *Для существования полного префиксного p-арного кода конечной мощности m необходимо и достаточно условие $p-1|m-1$.* (2) *Число всех полных префиксных p-арных кодов мощности $(p-1)k+1$ равно числу Фусса $F(p;k)$.*

Доказательство. (1) Необходимость. Если C – полный префиксный p-арный код, то, в силу лемм 2 и 4, дерево $T(C)$ полно и $m = |C|$ равно числу листьев в $T(C)$, которое по лемме 1(2) равно $(p-1)k+1$ для некоторого k, т. е. $p-1|m-1$.

Достаточность. Пусть $m = (p-1)+1$ для некоторого k. Построим полный префиксный код над A_p мощности m. Положим

$$K = \{e, a_1, a_1^2, \ldots, a_1^{k-1}\},\ B = \{a_2, \ldots, a_p\} \text{ и } C = \{a_1^k\} \cup K \cdot B.$$

Очевидно, C — префиксный код и $|C| = m$. Проверим условие полноты. Если $w \in A_p^*$, то $w \in a_1^l(BA_p)^n$ для некоторых $l, n \geqslant 0$. Если $l \geqslant k$, то $a_1^k \leqslant w$; если $l < k$, то $w \leqslant a_1^k$ при $n = 0$, и $a_1^l a_j \leqslant w$ при $n > 0$, где $a_1^l a_j \in KB$. Таким образом, существует полный префиксный p-арный код мощности m.

(2) вытекает из лемм 4 и 1(1). □

Следствие 6. *Полные префиксные бинарные коды конечной мощности m существуют при любом $m > 0$, и число таких кодов равно числу Каталана C_{m-1}.* □

Обозначим через $\mathcal{T}(p;k,m)$ множество всех p-арных k-деревьев с m листьями и положим $t_{km}(p) = |\mathcal{T}(p;k,m)|$. Тогда $t_{km}(p)$ — это также число всех префиксных p-арных кодов C мощности m, для которых $|T(C)| = k$. Пусть $t = t(p;x,y) = \sum_{k,m\geqslant 0} t_{km}(p)x^k y^m$ — производящая функция чисел $t_{km} = t_{km}(p)$. Следующий результат — основной в данной заметке.

Теорема 7. (1) *Степенной ряд $t = t(p;x,y)$ удовлетворяет уравнению $t = 1 + xy + x(t^p - 1)$.*

(2) *Для* $p, k, m > 0$

$$t_{km}(p) = \frac{1}{k}\binom{k}{m}\sum_{i\geqslant m}(-1)^{i-m}\binom{k-m}{i-m}\binom{p(k-i)}{k-1},$$

где верхний предел суммирования равен $\left\lfloor \frac{(p-1)k+1}{p} \right\rfloor$.

Доказательство. (1) Легко понять, что $t_{00} = t_{11} = 1$ (имеются единственные пустое и одноузловое деревья) и $t_{kk} = 0$ при $k > 1$ (все узлы — если их больше одного – не могут быть листьями), $t_{k0} = 0$ при $k > 0$ (у непустого дерева должны быть листья) и $t_{km} = 0$ при $k < m$.

Пусть $m(T)$ — число листьев дерева T. Рекурсивное представление $T \simeq (T_1, \ldots, T_p)$ при $|T| \geqslant 2$ даёт $|T| = 1 + \sum_{i=1}^{p} |T_i|$, $m(T) = \sum_{i=1}^{p} m(T_i)$ и, следовательно, при $k \geqslant 2$

$$\mathcal{T}(p; k, m) \simeq \bigcup_{\substack{k_1 + \ldots + k_p \,=\, k-1 \\ m_1 + \ldots + m_p \,=\, m}} \mathcal{T}(p; k_1, m_1) \times \ldots \times \mathcal{T}(p; k_p, m_p),$$

откуда

$$t_{km} = \sum_{\substack{k_1 + \ldots + k_p \,=\, k-1 \\ m_1 + \ldots + m_p \,=\, m}} t_{k_1 m_1} t_{k_2 m_2} \ldots t_{k_p m_p}.$$

Поэтому (заменяя ниже во второй сумме k на $l+1$)

$$t = \sum_{k<2 \vee m=0} t_{km} x^k y^m + \sum_{k\geqslant 2, m\geqslant 1} t_{km} x^k y^m =$$

$$= 1 + xy + x \sum_{l\geqslant 1, m\geqslant 1} \sum_{\substack{l_1 + \ldots + l_p \,=\, l \\ m_1 + \ldots + m_p \,=\, m}} t_{l_1 m_1} \ldots t_{l_p m_p} x^l y^m.$$

Если внутреннюю сумму справа обозначить через $S(l, m)$, то

$$\sum_{l\geqslant 1, m\geqslant 1} S(l, m) = t^p - S(0,0) - S(0,1) - S(1,0) = t^p - 1.$$

Таким образом, $t = 1 + xy + x(t^p - 1)$.

(2) Вводя ряд $f = t - 1$, получаем уравнение $f = x(y - 1 + (1+f)^p)$. Положим $R = \mathbb{Q}[[y]]$ (здесь $\mathbb{Q}$ — поле рациональных чисел). Тогда $f \in R[[x]]$ (в силу стандартного отождествления $\mathbb{Q}[[x, y]] = \mathbb{Q}[[y]][[x]]$). Если определить $\varphi(u) = y - 1 + (1+u)^p$, то $\varphi(u) \in R[u]$ и $f = x\varphi(f)$.

Применим формулу Лагранжа–Бюрмана: при $k>0$

$$\begin{aligned}[x^k]f &= \frac{1}{k}[u^{k-1}]\varphi^k = \frac{1}{k}[u^{k-1}](y-1+(1+u)^p)^k = \\ &= \frac{1}{k}[u^{k-1}]\sum_{i=0}^{k}\binom{k}{i}(y-1)^i\sum_{j=0}^{p(k-i)}\binom{p(k-i)}{j}u^j = \\ &= \frac{1}{k}\sum_{i=0}^{k}\binom{k}{i}\binom{p(k-i)}{k-1}(y-1)^i.\end{aligned}$$

Далее, так как $t=f+1$, для $k,m>0$

$$t_{km}(p) = [x^k y^m]t = [y^m][x^k]f = \frac{1}{k}\sum_{i=0}^{k}\binom{k}{i}\binom{p(k-i)}{k-1}[y^m](y-1)^i =$$

$$= \frac{1}{k}\sum_{i=0}^{k}\binom{k}{i}\binom{p(k-i)}{k-1}\binom{i}{m}(-1)^{i-m}.$$

Поскольку $\binom{k}{i}\binom{i}{m} = \binom{k}{m}\binom{k-m}{i-m}$,

$$t_{km}(p) = \frac{1}{k}\binom{k}{m}\sum_{i=0}^{k}(-1)^{i-m}\binom{k-m}{i-m}\binom{p(k-i)}{k-1}.$$

Из формулы видно, что индексу i достаточно пробегать отрезок от m до $\left\lfloor\frac{(p-1)k+1}{p}\right\rfloor$. □

Найденная формула явно проверена для $p=2,3$ и $k\leqslant 4$, а также для $p=4$ и $k\leqslant 3$.

Следствие 8. *Для любых $p,k\geqslant 0$ выполняется тождество*

$$\sum_{j=0}^{k}\binom{pj}{k}\binom{k+1}{j}\sum_{l=0}^{k-j}(-1)^l\binom{k+1-j}{l} = \binom{p(k+1)}{k}.$$

Доказательство. Случай $p=0\vee k=0$ проверяется непосредственно. Пусть $p,k>0$. Из равенства $\bigcup_{m\geqslant 1}\mathcal{T}(p;k,m) = \mathcal{T}(p;k)$, леммы 1 и теоремы 7 следует, что

$$\sum_{m=1}^{k}\binom{k}{m}\sum_{i=m}^{k}(-1)^{i-m}\binom{k-m}{i-m}\binom{p(k-i)}{k-1} = \binom{pk}{k-1}.$$

Меняя порядок суммирования и вводя новые индексы суммирования – сначала $j=k-i$, затем $l=k-j-m$, — получаем левую часть равенства в виде

$$\sum_{j=0}^{k-1}\binom{pj}{k-1}\sum_{l=0}^{k-j-1}(-1)^l\binom{l+j}{l}\binom{k}{k-j-l}.$$

Замечая, что $\binom{l+j}{l}\binom{k}{k-j-l}=\binom{k}{l+j}\binom{l+j}{j}=\binom{k}{j}\binom{k-j}{j}$, и заменяя k на $k+1$, приходим к требуемому тождеству. $\square$

Следствие 9. *Для любых $p,k>0$ выполняется тождество*

$$\sum_{j=0}^{k-1}(-1)^j\binom{k}{j}\binom{p(k-j)}{k}=p^k.$$

Доказательство. Число $t_{k+1,1}(p)$ всех префиксных p-арных кодов C мощности 1 с $|T(C)|=k+1$ — это просто число слов длины k в алфавите A_p, т. е. p^k. В равенстве $t_{k+1,1}(p)=p^k$ заменим индекс суммирования: $i=j+1$. $\square$

Следующее утверждение можно доказать индуктивно, но поучительно воспользоваться формулой из теоремы 7.

Следствие 10. *Пусть $p>0, n\geqslant 0$. Если $k=pn+1$ и $m=k-n$, то $\mathcal{CT}(p;n)=\mathcal{T}(p;k,m)$.*

Доказательство. Очевидно, $\mathcal{CT}(p;n)\subseteq\mathcal{T}(p;k,m)$. Чтобы доказать равенство, достаточно доказать равномощность этих множеств. Верхний предел суммирования в формуле для $t_{km}(p)$ равен $\lfloor((p-1)k+1)/p\rfloor=\lfloor k-(k-1)/p\rfloor==k-n=m$, а $\binom{p(k-m)}{k-1}=\binom{pn}{pn}=1$. Поэтому

$$|\mathcal{T}(p;k,m)|=t_{km}(p)=\frac{1}{k}\binom{k}{m}=\frac{1}{pn+1}\binom{pn+1}{n}=|\mathcal{T}(p;n)|.\quad\square$$

Для перечисления бинарных префиксных кодов имеется существенно более простая формула (см. ниже теорему 13), нежели даёт теорема 7. В [3] приведён (с ошибкой) сходный результат подсчёта полных бинарных деревьев с m парами «братьев–листьев» и указано, что эквивалентные утверждения получены в работах Избицки [4], Риордана [5] и Шапиро [6]. Здесь даётся короткое биективное доказательство. Но сначала приведём дополнительные сведения.

Пусть $X=\{x,\overline{x},l,r\}$ и морфизм $\delta: X^*\to(\mathbb{Z},+)$ задан равенствами $\delta(x)=1$, $\delta(\overline{x})=-1$, $\delta(l)=\delta(r)=0$. (Двухцветный) *язык Моцкина* $\mathcal{M}$ — это множество всех слов $w\in X^*$, удовлетворяющих условиям $\delta(w)=0$ и $(u\leqslant w\Rightarrow\delta(u)\geqslant 0)$. Взяв $Y=\{x,\overline{x}\}$, получим (ограниченный) *язык Дика* $\mathcal{D}=\mathcal{M}\cap Y^*$; таким образом, слова Дика получаются из слов Моцкина стиранием букв l и r. Положим $\mathcal{M}_n = \mathcal{M} \cap X^*$, $\mathcal{D}_n = \mathcal{D} \cap Y^{2n}$.

Введём ещё обозначения: $\mathcal{B}_k = \mathcal{T}(2;k)$, $\mathcal{CB}_k = \mathcal{CT}(2;k)$, $\mathcal{B}_k(m) = \mathcal{T}(2;k,m)$. Имеется биекция $\psi : \mathcal{B}_k \to \mathcal{M}_{k-1}$, определяемая следующим образом. Пометим рёбра дерева $T \in \mathcal{B}_k$ буквами из X: буквой l (соответственно, r) — ребро, ведущее к левому (соответственно, правому) сыну, не имеющему брата; буквами x и $\overline{x}$ — рёбра, ведущие, соответственно, к левому и правому братьям. Слово $\psi(T)$ получается сканированием дерева T сверху вниз (при обходе слева направо). То же правило задаёт биекцию $\mathcal{CB}_k \to \mathcal{D}_k$. Эти биекции, наряду с леммой 1, показывают, что выполняется

Лемма 11. $|\mathcal{D}_k| = |\mathcal{CB}_k| = |\mathcal{B}_k| = |\mathcal{M}_{k-1}| = C_k$. □

Лемма 12. *Число листьев в непустом конечном бинарном дереве на 1 больше числа узлов, имеющих двух сыновей.*

Доказательство — индукция с использованием рекурсивного представления деревьев. □

Теорема 13. *При $k, m > 0$ число $t_{km}(2) = |\mathcal{B}_k(m)|$ равно*

$$2^{k-2m+1}\binom{k-1}{2m-2}C_{m-1} = \frac{1}{k}2^{k-2m+1}\binom{k}{m}\binom{k-m}{m-1}.$$

Доказательство. По лемме 12 искомое число есть число бинарных k-деревьев с $m-1$ узлами, имеющими двух сыновей, и равно, в свою очередь, числу слов Моцкина длины $k-1$ с $m-1$ вхождениями буквы x. Чтобы сформировать такое слово, нужно из $k-1$ мест выбрать $2m-2$ мест для букв x и $\overline{x}$ ($\binom{k-1}{2m-2}$ способов), разместить там слово Дика (C_{m-1} способов), а оставшиеся места заполнить буквами l и r (2^{k-2m+1} способов). В итоге

$$|\mathcal{B}_k(m)| = 2^{k-2m+1}\binom{k-1}{2m-2}C_{m-1}.$$

Множитель при степени двойки равен

$$\frac{1}{m}\binom{k-1}{2m-2}\binom{2m-2}{m-1} = \frac{1}{k}\binom{k}{m}\binom{k-m}{m-1}. \quad \square$$

Заметим, что переход от равенства $\mathcal{B}_{k+1} = \bigcup_{m\geqslant 0}\mathcal{B}_{k+1}(m+1)$ к равенству мощностей даёт известное тождество Тушара

$$C_{k+1} = \sum_{m=0}^{\lfloor k/2 \rfloor} 2^{k-2m}\binom{k}{2m}C_m.$$

Записывая число $t_{km}(2)$ двумя способами (теоремы 7 и 13), получаем

Следствие 14. *Для любых $k, m > 0$ выполняется тождество*

$$\sum_{i=0}^{k}(-1)^{i-m}\binom{k-m}{i-m}\binom{2k-2i}{k-1} = 2^{k-2m+1}\binom{k-m}{m-1}. \quad \square$$

При перечислении префиксных кодов C параметр $k = |T(C)|$ выглядит не слишком естественным (хотя он естествен для деревьев). Но фиксация k обеспечивает конечность перечисляемых множеств (в случае полных префиксных кодов для этого достаточен параметр $m = |C|$). Другим параметром, обеспечивающем конечность множества кодов, может служить длина $\lambda(C)$ кода C, $\lambda(C) = \max\{|w| | w \in C\}$. Это также высота дерева $T(C)$, т. е. максимальная длина цепей, начинающихся в корне дерева. Я думаю, представляет интерес перечисление префиксных кодов и деревьев в зависимости от параметра λ.

Благодарности. Автор признателен В. А. Баранскому за предварительное обсуждение результатов и В. В. Расину за помощь с TEXом.

Список литературы

1. Коряков И. О. Сильно префиксные коды и обобщенная формула Рэйни / И. О. Коряков // Изв. Урал. гос. ун-та. – 2010. – №74. – С. 57–66.
2. Лаллеман Ж. Полугруппы и комбинаторные приложения / Ж. Лаллеман. – М. : Мир, 1985.
3. Shapiro L. W. Positive definite matrices and Catalan numbers / L. W. Shapiro // Proc. Amer. Math. Soc. – 1984. – Vol. 90, №3. – P. 488–496.
4. Izbicki H. Über Unterbäume eines Baumes / H. Izbicki // Monatsh. Math. – 1970. – Vol. 74. – P. 56–62.
5. Riordan J. A note on Catalan parentheses / J. Riordan // Amer. Math. Monthly. – 1973. – Vol. 80. – P. 904–906.
6. Shapiro L. W. A short proof of an identity of Touchard's concerning Catalan numbers / L. W. Shapiro // J. Combinatorial Theory. Ser. A. – 1976. – Vol. 20, №3. – P. 375–376.

I. O. Koryakov

Finite prefix codes and finatary trees

Abstract. We find an explicit formula for the number of all p-ary k-trees with m leaves. Analyzing this formula, we arrive at a number of combinatorial identities.

Keywords: tree; prefix code; generating function; Fuss number; Catalan number; Motzkin language.

Коряков Игорь Олегович, старший научный сотрудник, Институт математики и компьютерных наук, Уральский федеральный университет, 620000 Екатеринбург, пр. Ленина, 51; тел.: (343)3507579

Koryakov Igor O., senior scientist, Ural Federal University, Lenina 51, 620000 Ekaterinburg; Phone: (343)3507579