ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Январь—февраль 2014. Том 21, № 1. С. 3–14

УДК 519.114

О РАЗЛИЧЕНИИ СЛОВ ВХОЖДЕНИЯМИ ПОДСЛОВ *)

М. Н. Вялый, Р. А. Гимадеев

Аннотация. Получены нижние оценки сложности различения слов кратностями вхождений подслов с учётом позиции подслова в слове. Доказано, что в случае подслов длины 1 оценка оптимальна с точностью до мультипликативного множителя. Рассмотрена связь задачи различения слов вхождениями подслов с задачей различения слов автоматами.

Ключевые слова: подслово, различение слов, круговой многочлен, автомат.

Задачи о сложности различения слов хорошо известны в комбинаторике слов и рассматриваются в различных постановках.

Например, в [1] решена задача о различении слов подсловами. Обобщается эта задача двумя способами: (i) число вхождений подслова в слово заменяется числом вхождений подслова в слово на позициях, имеющих фиксированный остаток по некоторому модулю; (ii) вместо двоичных последовательностей рассматриваются возрастающие последовательности натуральных чисел. Различение происходит в два этапа: на первом выбирается некоторое число и последовательности редуцируются по модулю этого числа, а на втором последовательности остатков сравниваются по кратностям вхождения в них подслов заданной длины. В обоих случаях получаем степенны́е нижние оценки на сложность различения чисел.

Другая важная задача различения подслов — задача различения подслов автоматами. Она далека от решения: наилучшие известные оценки — нижняя оценка $\Omega(\log n)$ и верхняя оценка Робсона $O(n^{2/5}\log^{3/5} n)$ — различаются экспоненциально. Различение слов вхождениями подслов связано с этой задачей. Ниже обсуждается эта связь и указываются возможные подходы к улучшению оценки Робсона.

*)Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 11–01–00398, 12–01–00864), а также гранта президента РФ по поддержке ведущих научных школ (проект НШ–4652.2012.1).

1. Модулярный состав слова

Напомним, что *составом* (или *вектором Парика*) слова w в конечном алфавите $\Sigma = (\sigma_1, \ldots, \sigma_s)$ называется набор чисел $(n_1, \ldots, n_s)$, где n_i — количество вхождений символа σ_i в слово w.

Очевидно, что состав слова различает лишь слова длины 1: у слов 01 и 10 в двоичном алфавите состав одинаков.

Более точное описание слова получается, если указать частичную информацию о позициях символов в слове.

Обозначим через $n(\sigma_i, w; a, q)$ количество вхождений символа σ_i на позициях с номерами, которые имеют остаток a по модулю q. Более формально, для слова $w = w_0 w_1 \ldots w_{n-1}$ по определению выполняется равенство $n(\sigma_i, w; a, q) = |\{j \mid w_j = \sigma_i \text{ и } j \equiv a \pmod q\}|$.

Назовём *ℓ-модулярным составом* слова w совокупность кратностей вхождения символов на указанных позициях, т. е. чисел $n(\sigma_i, w; a, q)$ для всех $1 \leqslant i \leqslant s$, $0 \leqslant a < q \leqslant \ell$.

Определим $M_s(n)$ как наименьшее ℓ такое, что любая пара различных слов v, w длины не больше n в алфавите из s символов различается ℓ-модулярным составом.

Сложность различения модулярным составом почти не зависит от размера алфавита. По определению $M_1(n) = 1$: слова разной длины различаются уже по модулю 1. Поэтому далее рассматриваем только различение слов одинаковой длины.

Утверждение 1. *При $s \geqslant 2$ имеет место равенство $M_s(n) = M_2(n)$.*

Доказательство. Неравенство $M_s(n) \geqslant M_2(n)$ следует из того, что двоичные слова можно рассматривать как слова в любом большем алфавите.

Докажем неравенство в обратную сторону. Пусть $u \neq v$ — пара слов в алфавите из $s \geqslant 2$ символов с одинаковым ℓ-модулярным составом, причём $u_i \neq v_i$. Обозначим через ψ морфизм, который задаётся следующим образом: $\psi(u_i) = 1$, $\psi(\sigma) = 0$ при $\sigma \neq u_i$. Тогда $\psi(u)$, $\psi(v)$ — пара различных двоичных слов с одинаковым ℓ-модулярным составом, так как $\psi(u)_i \neq \psi(v)_i$. Утверждение 1 доказано.

В силу утверждения 1 в дальнейшем будем опускать индекс, обозначать сложность различения модулярным составом через $M(n)$ и считать, что в алфавите есть по крайней мере два различных символа.

Сложность различения слов модулярным составом оценим с точностью до полилогарифмического множителя.

Лемма 1. $M(n) = \Omega(n^{1/2} \log^{-1/2} n)$.

Доказательство. Кратность вхождения любого символа не превосходит длины слова n. Остатков по модулю m ровно m штук, запись всех этих остатков требует $O(m \log n)$ битов. Поэтому для записи ℓ-модулярного состава требуется $O(\ell^2 \log n)$ битов. Поскольку всего двоичных слов 2^n, получаем требуемую оценку. Лемма 1 доказана.

Для доказательства верхней оценки на сложность различения модулярным составом потребуется алгебраическое описание модулярного состава.

Для двоичного слова w рассмотрим производящую функцию единиц $W(t) = \sum\limits_j t^{n_j}$, где n_j — номер позиции j-й единицы в слове w. Аналогично для двоичного слова u определяется производящая функция единиц $U(t)$. По определению $W(t)$ — многочлен степени не выше n.

Лемма 2. *Если двоичные слова u и w не различаются кратностями вхождения единиц по модулю m, то $U(t) - W(t)$ делится на круговой многочлен $\Phi_m(t)$.*

Доказательство. Проверим, что $U(t) - W(t)$ делится на многочлен $t^m - 1$, кратный $\Phi_m(t)$. Действительно, по определению модулярного состава в $U(t)$ входит $n(1, u; a, m)$ мономов вида t^{a+km}, а в $W(t)$ таких мономов $n(1, w; a, m) = n(1, u; a, m)$. Разбивая мономы на пары для каждого a, получаем представление $U(t) - W(t)$ в виде суммы многочленов вида $t^{a+k'm} - t^{a+k''m}$, делящихся на $t^m - 1$. Лемма 2 доказана.

Теорема 1. $M(n) = O(n^{1/2})$.

Доказательство. Круговые многочлены взаимно просты. Поэтому из того, что ненулевой многочлен $U(t) - W(t)$ степени не выше n делится на произведение круговых $\Phi_m(t)$, $1 \leqslant m \leqslant \ell$, следует оценка

$$n \geqslant \sum_m^{\ell} \deg \Phi_m(t).$$

Как известно, $\deg \Phi_m(t) = \varphi(m)$, где $\varphi(n)$ — функция Эйлера. Из теоремы Мертенса [2, с. 82] имеем

$$\sum_{m=1}^{\ell} \varphi(m) = \frac{3}{\pi^2}\ell^2 + O(\ell \log \ell),$$

откуда получаем оценку теоремы. Теорема 1 доказана.

2. Модулярный подсловный состав

По определению слово $u = u_0u_1 \dots u_{\ell-1}$ входит как *подслово* в слово $w = w_0w_1 \dots w_{n-1}$ на позиции i, если для любого $0 \leqslant j < \ell$ выполняется равенство $w_{i+j} = u_j$. Как видно из этого определения, позиции в словах нумеруются, начиная с 0.

Число вхождений слова u в слово w обозначим через $n(u, v)$. *Подсловным составом длины* ℓ слова w называется набор кратностей вхождения всех подслов длины не больше ℓ в слово w.

Слова, не различающиеся подсловным составом длины k, называются *k-абелево эквивалентными*. k-Абелева эквивалентность изучалась в ряде работ по комбинаторике слов, например, в [4], где рассматриваются свойства k-абелевых периодов.

Для задач различения слов k-абелева эквивалентность слишком слаба. В [1] показано, что существуют слова длины n, которые не различаются подсловным составом длины вплоть до $n/2 - 1$.

Поэтому рассмотрим более детальное описание вхождений подслов, которое учитывает номера позиций, на которых входит данное подслово.

Обозначим через $n(u, w; a)$ количество подслов u, входящих в слово w на позициях с номерами $a \pmod{\ell}$, где ℓ — длина слова u. Числа $n(u, w; a)$ будем называть *кратностями вхождения выровненных слов*. Кратности вхождения выровненных подслов длины не больше ℓ образуют *ℓ-модулярный подсловный состав* слова.

Как и выше, нас интересует сложность различения слов модулярным подсловным составом. Определим $L_s(n)$ как наименьшее ℓ такое, что для любой пары различных слов v, w длины n в алфавите из s символов различаются наборы кратностей вхождения $n(u, v; a)$ и $n(u, w; a)$, где $|u| = q \leqslant \ell$, $0 \leqslant a < q$.

Замечание. В определении модулярного подсловного состава мы рассматриваем вхождения слов на позициях, номера которых имеют заданный остаток по модулю длины слова. Однако нетрудно видеть, что количество ℓ-модулярных вхождений слова длины $\ell' < \ell$ выражается через компоненты ℓ-модулярного подсловного состава. В частности, при $\ell' = 1$ получаем, что компоненты ℓ-модулярного состава слова выражаются через компоненты ℓ-модулярного подсловного состава слова. Поэтому $L_s(n) \leqslant M(n)$.

Аналогично утверждению 1 проверяется

Утверждение 2. *При $s \geqslant 2$ выполняется $L_s(n) = L_2(n)$.*

Доказательство. Неравенство $L_s(n) \geqslant L_2(n)$ следует из того, что двоичные слова можно рассматривать как слова в любом большем алфавите.

Докажем неравенство в обратную сторону. Пусть $u \neq v$ — пара слов в алфавите из $s \geqslant 2$ символов с одинаковым ℓ-модулярным подсловным составом, причём $u_i \neq v_i$. Как и при доказательстве утверждения 1, рассмотрим морфизм ψ такой, что $\psi(u_i) = 1$, $\psi(\sigma) = 0$ при $\sigma \neq u_i$. Тогда $\psi(u)$, $\psi(v)$ — пара различных двоичных слов с одинаковым ℓ-модулярным подсловным составом, так как $\psi(u)_i \neq \psi(v)_i$. Утверждение 2 доказано.

Учитывая утверждение 2, будем опускать указание на размер алфавита в сложности различения модулярным подсловным составом.

Теорема 2. $L(n) = \Omega(n^{1/3} \log^{-1/3} n)$.

Доказательство. Пусть u, w — два различных двоичных слова длины $C\ell^2 \log \ell$, которые не различаются ℓ-модулярным составом, где C — некоторая константа. Существование таких слов гарантирует лемма 1.

Выберем простое число $\ell < p \leqslant 2\ell$ (такое число найдётся в силу постулата Бертрана [2]). Обозначим через $p \cdot u$, $p \cdot w$ слова, которые получаются растяжением слов u и w в p раз. Более точно эта операция описывается следующим образом: длина слова $p \cdot w$ в p раз больше длины слова w; на позиции pj в слове $p \cdot w$ стоит единица тогда и только тогда, когда в слове w единица стоит на позиции j, на остальных позициях стоят нули. (Напомним, что позиции в слове нумеруем, начиная с 0.) Например,

$$3 \cdot 1011 = 100000100100.$$

Докажем, что слова $p \cdot u$, $p \cdot w$ не различаются ℓ-модулярным подсловным составом, откуда и будет следовать оценка теоремы.

По построению в слова $p \cdot u$, $p \cdot w$ входят только подслова длины $\ell' \leqslant \ell$ вида $0^{\ell'}$ и $0^a 10^{\ell'-1-a}$, где $0 \leqslant a \leqslant \ell' - 1$.

Подслово $0^a 10^{q-1-a}$ входит в слово $p \cdot w$ на позиции j тогда и только тогда, когда $j + a = pj'$ и на позиции j' в слове w стоит единица. Таким образом, при $q \leqslant \ell$ получаем соотношение $j' \equiv p^{-1}(j + a) \pmod q$, из которого следует равенство

$$n(0^a 10^{q-1-a}, w; b) = n(1, w; p^{-1}(a + b) \bmod q, q). \tag{1}$$

Из (1) и того, что u и w не различаются ℓ-модулярным составом, следует, что $p \cdot u$, $p \cdot w$ не различаются ℓ-модулярным подсловным составом. Действительно, количество вхождений подслов вида 0^r выражается через количество вхождений остальных подслов. Теорема 2 доказана.

3. Различение числовых последовательностей

Рассмотрим другое обобщение модулярного состава двоичного слова и задачу различения числовых последовательностей.

Основное соответствие между (двоичными) словами и числовыми последовательностями задаётся следующим образом: слову w сопоставляем последовательность $x_w = (x_i)$ номеров позиций, на которых в слове w стоят единицы. Таким образом, слову длины n сопоставляется монотонно возрастающая последовательность натуральных чисел, которые не превосходят n. Например, слову 101100 сопоставляется последовательность $(0, 2, 3)$.

Отображение $w \mapsto x_w$ не инъективно, поскольку теряется информация о конечных нулях слова. В дальнейшем считаем, что граница слова задана явно, и рассматриваем возрастающие последовательности, в которых все члены не превосходят n. На таких последовательностях отображение становится взаимно однозначным. В частности, этих последовательностей ровно 2^n штук (включая пустую). Будем обозначать множество последовательностей указанного вида через B_n.

Пусть m — целое положительное число. Тогда $x \bmod m$ — последовательность $(x_i \bmod m)$ остатков от деления членов последовательности на m. Эту последовательность можно также рассматривать как слово в алфавите $[m] = \{0, 1, \ldots, m-1\}$ из m символов.

При доказательстве теоремы 1 уже фактически использовано представление ℓ-модулярного состава в виде списка составов последовательностей $x_w \bmod m$, где $1 \leqslant m \leqslant \ell$.

Рассмотрим подсловный состав последовательностей $x_w \bmod m$ и сложность различения числовых последовательностей подсловным составом. Обозначим через $Q(\ell, n)$ наименьшее q такое, что любая пара x, y различных последовательностей из B_n различается подсловным составом последовательностей $x \bmod m$ и $y \bmod m$ для некоторого $m \leqslant q$.

Из определений следует равенство $Q(1, n) = M(n)$. Поэтому из леммы 1 получаем нижнюю оценку

$$Q(1, n) = \Omega(n^{1/2} \log^{-1/2} n).$$

Докажем аналогичную оценку для подслов длины 2.

Лемма 3. $Q(2, n) = \Omega(n^{1/2} \log^{-1/2} n)$.

Доказательство. Рассмотрим множество $F_n \subset B_n$ таких последовательностей, в которых расстояние между соседними членами принимает ровно два значения: 1 или 2. Количество таких последовательностей

равно n-му числу Фибоначчи и растёт как φ^n, где $\varphi = (1+\sqrt{5})/2 > 1$ — золотое сечение.

Если $x \in F_n$, то в последовательности $x \bmod m$ подслова длины 2 имеют вид $(a, a+1)$ или $(a, a+2)$ (сложение по модулю m). Поэтому для записи кратностей вхождения подслов длины 2 в этом случае нужно $2m \log n$ битов. Далее рассуждаем аналогично доказательству леммы 1. Лемма 3 доказана.

Для подслов произвольной длины получаем следующую оценку.

Теорема 3. $Q(\ell, n) = \Omega(n^{1/2} \log^{-1/2} n\ell^{-1/2})$.

Доказательство. Выберем последовательности x' и y' из $B_{n'}$, $n' = q^2 \log q$, которые не различаются подсловными составами длины, меньшей или равной 2, вплоть до модуля q. Лемма 3 гарантирует существование таких последовательностей.

Умножим все члены последовательности на $\ell + 1$ и заменим каждый член $(\ell+1)x'_j$ последовательностью длины $\ell+1$ вида

$$(\ell+1)x'_j, (\ell+1)x'_j + 1, (\ell+1)x'_j + 2, \ldots, (\ell+1)x'_j + \ell. \tag{2}$$

Полученные последовательности x, y принадлежат $B_{n'(\ell+1)}$.

Посчитаем кратность вхождения слова $w \in [m]^\ell$ в последовательность x. Любое подслово длины ℓ в этой последовательности либо имеет вид

$$((\ell+1)x'_{j-1} + \ell - k) \ldots ((\ell+1)x'_{j-1} + \ell)((\ell+1)x'_j) \ldots ((\ell+1)x'_j + \ell - k - 2),$$

либо является вставкой (2), отвечающей одному индексу j. Подслов второго типа имеется две разновидности: одни начинаются с $(\ell+1)x'_j$, а другие нет.

Число подслов первого типа определяется количеством соответствующих подслов длины 2 в последовательности x', а второго типа — модулярным составом (подсловами длины 1). Значит, построенные последовательности x и y имеют одинаковый подсловный состав длины ℓ по модулям вплоть до q. Отсюда получаем оценку теоремы. Теорема 3 доказана.

Изучено два разных способа обобщения подсловного состава слова. Разумеется, их можно скомбинировать. Обозначим через $\widehat{Q}(\ell, n)$ наименьшее q такое, что любая пара x, y различных последовательностей из B_n различается ℓ-модулярным подсловным составом последовательностей $x \bmod m$ и $y \bmod m$ для некоторого $m \leqslant q$.

Для $\widehat{Q}(\ell, n)$ справедлива нижняя оценка $\Omega(n^{1/3}\log^{-1/3} n\ell^{-1/3})$, аналогичная оценкам из теорем 2 и 3. Конструкция из доказательства теоремы 3 позволяет оценить $\widehat{Q}(\ell, n)$ через сложность различения подслов модулярными вхождениями подслов длины 2, для которых аналогично лемме 1 доказывается оценка $\Omega(n^{1/3}\log^{-1/3} n)$.

4. Задача различения слов автоматами

Обсудим связь различения слов вхождениями подслов с задачей различения слов автоматами.

Напомним, что детерминированный конечный автомат A задаётся набором $(Q, \Sigma, q_0, Q_a, \delta)$, где Σ — конечное множество (алфавит), Q — конечное множество (множество состояний), $q_0 \in Q$ — выделенное начальное состояние, $Q_a \subseteq Q$ — множество принимающих состояний, а δ — функция переходов $\delta\colon Q \times \Sigma \to Q$. Функция переходов продолжается на всё множество Σ^* слов в алфавите Σ по естественному индуктивному правилу

$$\delta(q, u_0 \dots u_\ell) = \delta(\delta(q, u_0 \dots u_{\ell-1}), u_\ell).$$

Автомат $A = (Q, \Sigma, q_0, Q_a, \delta)$ принимает слово u, если $\delta(q_0, u) \in Q_a$.

Слова v и w различаются автоматом A, если автомат принимает ровно одно из них. Обозначим через $\mathrm{sep}(v, w)$ наименьшее число состояний в автомате, различающем слова v и w. Сложность различения слов автоматами $S_k(n)$ равна максимуму $\mathrm{sep}(v, w)$ по всем парам различных слов длины не больше n в алфавите из k символов.

Зависимость от алфавита несущественна, если в алфавите есть хотя бы два символа.

Утверждение 3 [3]. *При $k \geqslant 2$ выполняется $S_k(n) = S_2(n)$.*

В силу этого наблюдения далее обозначаем сложность различения слов автоматами через $S(n)$.

Задача различения слов автоматами состоит в получении как можно более точных оценок на $S(n)$. Приведём лучшие из известных оценок.

Оценка $S(n) = \Omega(\log n)$ — наилучшая нижняя оценка. Она легко получается для пар слов разной длины и воспроизводится во многих работах. В [3] такая же оценка доказана и для различения пар слов одинаковой длины.

Рекордная верхняя оценка $S(n) = O(n^{2/5}\log^{3/5} n)$ получена Робсоном [5]. Конструкция различающих автоматов и доказательство корректности в этом случае весьма трудны.

Опишем связь между различением модулярными составами и различением автоматами.

Лемма 4. $S(n) = O(L(n) \log n)$.

Доказательство. Нам потребуется автомат, который подсчитывает кратность вхождения выровненных подслов. Более точно, автомат вычисляет $n(u, w; a)$ по модулю некоторого числа q и имеет $O(\ell q)$ состояний. При этом считаем, что $|u| = \ell$, а $u, w \in \{0,1\}^*$.

Кратность вхождений подслова в слово длины n не превосходит n. Любые два различных числа, не превосходящих n, различаются по некоторому модулю $O(\log n)$. Отсюда следует утверждение леммы.

Требуемый в приведённом выше рассуждении автомат является композицией двух автоматов: счётчика по модулю q (с q состояниями) и управляющего этим счётчиком автомата A, который ищет выровненные вхождения подслова. Автомат A имеет $a+2\ell$ состояний. Первые a состояний используются на начальном отрезке, обозначим их через $q_0, \ldots, q_{a-1}$. Остальные 2ℓ состояний индексируются парами (i, α), где $i \in \mathbb{Z}/\ell\mathbb{Z}$, $\alpha \in \{0,1\}$. Они служат для обнаружения вхождений подслова u на позициях с номерами $a \pmod{\ell}$.

Первые компоненты состояния автомата A меняются циклически по модулю ℓ так, что после завершения чтения начального отрезка автомат читает символ на позиции $j + k\ell$, находясь в состоянии (j, α).

Вторая компонента состояния автомата равна 1 тогда и только тогда, когда на очередном цикле длины ℓ прочитан префикс слова u. Это условие легко формулируется в терминах таблицы переходов. Пусть $u = u_0 \ldots u_{\ell-1}$. Тогда

$$\delta((j,1),\sigma) = \begin{cases} (j+1,0), & \text{если } \sigma \neq u_j \text{ и } j \neq \ell-1, \\ (j+1,1) & \text{в противном случае,} \end{cases}$$

$$\delta((j,0),\sigma) = \begin{cases} (0,1), & \text{если } j = \ell-1, \\ (j+1,0) & \text{в противном случае.} \end{cases}$$

Сложение здесь подразумевается по модулю ℓ.

Осталось указать, как автомат A управляет подчинённым ему счётчиком по модулю q. Значение счётчика увеличивается на 1 при чтении символа $u_{\ell-1}$, если управляющий автомат A находится при этом в состоянии $(\ell-1, 1)$.

Из построения следует, что по окончании чтения слова подчинённый счётчик находится в состоянии t, которое по модулю q равно числу вхождений подслова u на позициях $a \pmod{\ell}$. Лемма 4 доказана.

Лемма 5. $S(n) = O(\widehat{Q}(\ell, n)\ell \log n)$.

Доказательство аналогично доказательству леммы 4. Легче всего представить искомый автомат как композицию автоматного преобразователя с $q = Q(\ell, n)$ состояниями, который по слову w строит слово $x_w \bmod \ell$, а на этом слове работает автомат из леммы 4. Лемма 5 доказана.

Из полученных нижних оценок можно заметить, что при использовании различения числовых последовательностей в задаче различения слов автоматами оптимален случай $\ell = O(1)$. На самом деле различение слов модулярным составом, т. е. при $\ell = 1$, возможно обратимыми автоматами с таким же числом состояний, что и в лемме 5. Для обратимых автоматов наилучшая оценка числа состояний, гарантирующих различение, также получена Робсоном в [6]. Эта оценка $O(\sqrt{n})$ на логарифмический множитель лучше оценки, получающейся из леммы 5 и теоремы 1. Однако эта разница несущественна и легко показать, что логарифмический множитель устраняется, если ограничиться лишь чётностью числа вхождений подслов (тогда подчинённый счётчик имеет $O(1)$ состояний). Возможность ограничиться чётностью следует из того, что алгебраическое доказательство теоремы 1 проходит и в том случае, когда многочлены рассматриваются над полем из двух элементов.

Обсудим теперь возможности улучшения оценки Робсона, основанные на использованном здесь подходе.

Из анализа приведённых выше конструкций видно, что для улучшени оценки Робсона достаточно показать точность оценки теоремы 2 на различение слов ℓ-модулярным подсловным составом (хотя бы с точностью до полилогарифмических множителей). С другой стороны, достаточно также доказать точность оценки сложности различения числовых последовательностей модулярными вхождениями подслов длины 2.

В [6] Робсон предложил различать слова автоматами $M_{x,y,d}$, которые принимают слова, содержащие нечётное количество единиц в позициях, номера которых дают остаток y по модулю d и которым предшествует нечётное количество единиц в позициях, номера которых дают остаток x по модулю d. Такие автоматы имеют $4d$ состояний. Робсон предположил, что они позволяют понизить оценку сложности распознавания обратимыми автоматами до $O(n^{1/3})$.

Во всех трёх случаях можно выразить требуемые оценки в терминах производящих функций. Самым простым и интересным является случай различения ℓ-модулярным составом.

Обозначим множество многочленов степени не выше m, которые делятся на $\prod\limits_{j \leqslant \ell} \Phi_j(t)$, через $I(\ell, m)$. Вопрос о точности оценки теоремы 2

оказывается связан с оценками «кодового расстояния» в пространстве многочленов $I(\ell, m)$, т. е. минимального размера носителя многочлена из $I(\ell, m)$ (число ненулевых мономов обозначим $d(\ell, m)$).

Оказывается, что из оценки

$$d(\ell, \ell^3 \mathrm{poly}(\log n)) = \Omega(\ell^2)/\mathrm{poly}(\log n)$$

следует, что $S(n) = O(n^{1/3})\mathrm{poly}(\log n)$. Это проверяется следующим рассуждением.

Пусть u, w в первый раз различаются в позиции $i > n^{1/3}$ (в противном случае слова, очевидно, различаются автоматами с $n^{1/3}$ состояниями). Обозначим через p_j подслово длины $j < \ell$, предшествующее позиции i. Тогда позиции, на которых любое из слов $p_j 0$ и $p_j 1$ входит в слова u и w, различаются. Выберем из множества этих слов префиксный код мощности ℓ (ни одно слово не является префиксом другого). Позиции вхождений подслов из префиксного кода в каждом из слов дизъюнктны. Хотя бы одно из этих слов входит не более n/ℓ раз. Применяя оценку для кодового расстояния $d(\ell, \ell^3 \mathrm{poly}(\log n))$, получаем верхнюю оценку на сложность различения модулярным составом и, как следствие, верхнюю оценку на сложность различения автоматами.

Вопрос об оценке кодового расстояния $d(\ell, m)$ остаётся открытым и, скорее всего, весьма труден.

В двух остальных случаях возникают производящие функции от двух переменных t_1, t_2 и оценка сводится к вопросу о принадлежности идеалу J_ℓ, порождённому многочленами $t_1^a - 1$, $t_2^a - 1$, $1 \leqslant a \leqslant \ell$.

Для различения модулярными подсловами длины 2 искомые производящие функции для последовательностей x_j имеют вид

$$\sum_j t_1^{N_j} t_2^{x_j}, \quad \text{где } N_j = \sum_{i<j} x_i. \tag{3}$$

Для автоматов Робсона производящие функции имеют вид

$$\sum_{i<j} t_1^{x_i} t_2^{x_j}. \tag{4}$$

Вопрос получения оценок различения слов автоматами из достаточных условий невхождения многочленов (3) или (4) в идеал J_ℓ пока остаётся открытым.

Хотя изложенные выше подходы к улучшению оценок Робсона не приводят к немедленному успеху, нам представляется, что они заслуживают дальнейшего изучения.

ЛИТЕРАТУРА

1. **Леонтьев В. К., Хошманд Асл М. Р.** Характеризация бинарных слов подсловами // Дискрет. анализ и исслед. операций. Сер. 1. — 2006. — Т. 13, № 1. — С. 65–76.
2. **Чандрасекхаран К.** Введение в аналитическую теорию чисел. — М.: Мир, 1974. — 188 с.
3. **Demaine E. D., Eisenstat S., Shallit J., Wilson D. A.** Remarks on separating words // Descriptional complexity of formal systems. — 2011. — P. 147–157. (Lect. Notes Comput. Sci.; Vol. 6808).
4. **Karhumaki J., Puzynina S., Saarela A.** Fine and Wilf's theorem for k-Abelian periods // Developments in language theory. — 2012. — P. 296–307. (Lect. Notes Comput. Sci.; Vol. 7410).
5. **Robson J. M.** Separating strings with small automata // Inf. Process. Lett. — 1989. — Vol. 30. — P. 209–214.
6. **Robson J. M.** Separating words with machines and groups // Inform. Théor. Appl. — 1996. — Vol. 30. — P. 81–86.

Вялый Михаил Николаевич,
e-mail: vyalyi@gmail.com
Гимадеев Ренат Айратович,
e-mail: renat.ariacas@gmail.com

Статья поступила
11 апреля 2013 г.

Переработанный вариант —
2 июля 2013 г.