ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Октябрь—декабрь 2003. Серия 1. Том 10, № 4, 70–78

УДК 519.714

О ПОРОЖДЕНИИ СЛОВ С ИСПОЛЬЗОВАНИЕМ ОПЕРАЦИИ КОМПОЗИЦИИ*)

Ю. В. Мерекин

Рассмотрено порождение слов с использованием операции композиции (впервые определенной А. И. Ширшовым). Доказано, что в классе схем композиции слов для получения нижних оценок сложности порождаемых слов применим суффиксный метод, который ранее использовался для схем конкатенации слов. Для коммутативных мономов при некоторых ограничениях на операцию композиции получено точное значение сложности их вычисления.

Введение

В настоящей статье продолжается исследование сложности процедуры построения слов, когда разрешается многократное использование уже построенных слов. Используя известную операцию композиции слов [7], которая является обобщением операции конкатенации, мы приводим определение схемы композиции произвольного слова Z в алфавите $\Sigma = \{a_1, \ldots, a_q\}$, $q \geqslant 1$, и вводим меру аддитивной сложности порождения слов. В дальнейшем изложении под сложностью везде понимается аддитивная сложность. Доказывается, что для получения нижней оценки сложности слов можно использовать суффиксный метод, который применялся ранее в работах автора (см., например, [3–6]).

В работе [1] исследовалась функция Шеннона для схем из функциональных элементов, реализующих вычисления в конечных абелевых группах. Ниже исследуется сложность процедуры порождения коммутативных мономов. В алгебре полиномов выражение $a_1^{k_1} \ldots a_q^{k_q}$, $k_i \geqslant 0$, $1 \leqslant i \leqslant q$, в алфавите $\Sigma = \{a_1, \ldots, a_q\}$, $q \geqslant 1$, называется *коммутативным мономом*. В дальнейшем изложении мы имеем дело только с такими мономами, поэтому слово "коммутативный"будем опускать.

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект 02-01-00939) и гранта Минобразования РФ Е02–6.0–250.

Коммутативная конкатенация (умножение) мономов определяется как $a_1^{k_1} \dots a_q^{k_q} \cdot a_1^{s_1} \dots a_q^{s_q} = a_1^{k_1+s_1} \dots a_q^{k_q+s_q}$. Операцию композиции, предложенную А. И. Ширшовым [7] для слов, распространим на мономы. Для мономов при некоторых ограничениях на операцию композиции получим точное значение сложности их вычисления.

Длиной $|W|$ слова W называется число символов в W. Слово V называется *подсловом* слова W и обозначается через $V \sqsubseteq W$, если для некоторых (возможно, пустых) слов X и Y справедливо равенство $W = XVY$. При пустом X подслово V называется *префиксом*, а при пустом Y — *суффиксом* слова W.

Для построения слов в алфавите Σ используем операцию композиции [7]. Пусть U_1, V_1, R — произвольные слова (возможно, R пусто), $U = U_1R$ и $V = RV_1$. *Композицией* слов U и V относительно слова R является слово U_1RV_1 и обозначается через $(U, V)_R = U_1RV_1$. При пустом R композиция совпадает с *конкатенацией* $U \bullet V$.

Последовательность слов $a_1, \dots, a_q, X, Y, \dots, Z$, которая обозначается через S, назовем *схемой композиции* слова Z, если для любого слова W из S, начиная со слова X, в S имеются такие слова $U = U_1R$ и $V = RV_1$ (возможно, R пусто, $U = V$), предшествующие слову W, что $W = (U, V)_R$. Если для построения всех слов $X, Y, \dots, Z$ применяется только операция конкатенации, то схема называется *схемой конкатенации* слова Z (см., например, [2, 3]). Удалим из S некоторые слова. В результате получим последовательность S_1. Если S_1 является схемой композиции, то она называется *подсхемой* схемы S и обозначается через $S_1 \subset S$.

Под *сложностью* $L_{sh}(S)$ схемы S композиции слова Z понимается число слов в последовательности $X, Y, \dots, Z$. Пусть $L_{sh}(Z) = \min L_{sh}(S)$, где минимум берется по всем схемам композиции слова Z. Величину $L_{sh}(Z)$ назовем *сложностью* слова Z. Схема S называется *оптимальной*, если $L_{sh}(Z) = L_{sh}(S)$. Через $L(Z)$ обозначается сложность слова Z для схем конкатенации.

При получении нижних оценок сложности слов в классе схем композиции слов применим суффиксный метод, предложенный в [3] для схем конкатенации слов. Суффиксный метод использует следующие представления слов. Пусть слово W представлено в виде $W = UxV$, где $x \in \Sigma$ и U, возможно, пусто. Если V является либо символом, отсутствующим в слове Ux, либо $V \sqsubseteq Ux$ и $xV \not\sqsubseteq U$, то слово V называется *максимальным суффиксом* слова W (однобуквенное слово по определению является своим максимальным суффиксом). Представление слова W в виде

$W = X_1 \bullet \ldots \bullet X_r$ называется *суффиксным представлением* , если длина слова X_1 равна единице, а каждое слово X_i, $1 \leqslant i \leqslant r$, является максимальным суффиксом слова $X_1 \bullet \ldots \bullet X_i$. Очевидно, что суффиксное представление любого слова единственно. Число операций конкатенации в суффиксном представлении слова W называется *суффиксной сложностью* слова W и обозначается через $L^*(W)$.

При нахождении суффиксного представления слова W могут возникнуть трудности с выделением максимальных суффиксов. Если в процессе построения суффиксного представления допустить замену максимального суффикса на суффикс большей длины, то можно получить некоторое разложение слова W, в котором число суффиксов не превосходит их числа в суффиксном представлении и, следовательно, может использоваться при получении нижней оценки для $L^*(W)$.

Пусть слово W представлено в виде $W = UV$, где $|V| > 1$. Если $V \not\sqsubseteq U$, то слово V называется *расширенным суффиксом* слова W (при пустом U слово V по определению является расширенным суффиксом). Представление слова W в виде $W = Y_1 \bullet \ldots \bullet Y_u$ называется *расширенным суффиксным представлением*, если

(i) каждый суффикс $Y_1, \ldots, Y_u$ является либо максимальным, либо расширенным;

(ii) среди $Y_1, \ldots, Y_u$ имеется хотя бы один расширенный суффикс.

Расширенное суффиксное представление слова, вообще говоря, не единственно. Число операций конкатенации в i-м расширенном представлении (предполагается, что все такие представления как-то перенумерованы) называется *сложностью i-го расширенного суффиксного представления* и обозначается через $L_i^{**}(W)$. Справедливы

Предложение 1. *Для любых слов U и V выполняется неравенство $L^*(UV) \geqslant L^*(U)$.*

Предложение 2. *Для произвольного слова W и любого его i-го расширенного суффиксного представления выполняется неравенство*

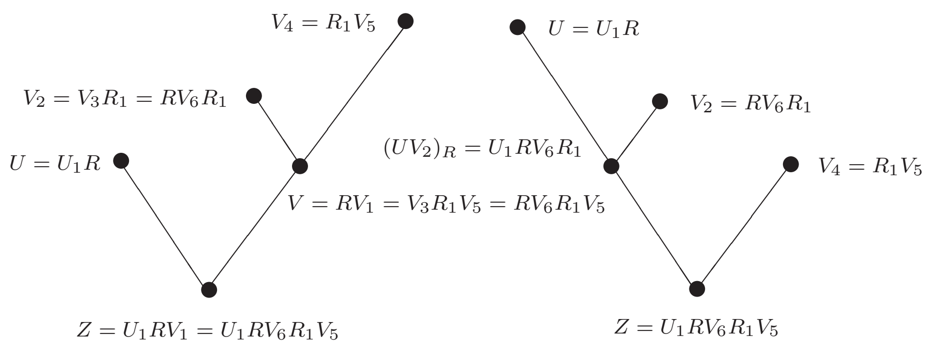
$$L^*(W) \geqslant L_i^{**}(W).$$

§ 1. Суффиксный метод получения нижних оценок сложности схем композиции слов

В [3] получено неравенство $L(W) \geqslant L^*(W)$. Основным результатом настоящего параграфа является утверждение о том, что $L_{sh}(W) \geqslant L^*(W)$. В классе схем конкатенации слов для получения нижних оценок сложности использовалось суффиксное представление слов. В доказательствах ключевую роль сыграло свойство соотношения слов в заключительной операции оптимальной схемы конкатенации слова. Следующая лемма обобщает это свойство для класса схем композиции слов.

Лемма. *Для любого слова Z существует оптимальная схема композиции, содержащая такие слова $U = U_1R$ и $V = RV_1$, что $Z = (U,V)_R$ (возможно, R пусто), и слово V является либо подсловом слова U, либо символом, отсутствующим в слове U.*

Доказательство. В случае, когда слово V является символом, отсутствующим в слове U, утверждение леммы очевидно. Докажем лемму при $V \sqsubseteq U$. Пусть S — некоторая оптимальная схема композиции слова Z, содержащая такие слова $U = U_1R$ и $V = RV_1$, что $Z = (U,V)_R = U_1RV_1$ и слово V, $|V| > 1$, не является подсловом слова U. В этом случае подсхема $S_1, S_1 \subset S$, композиции слова U не содержит слова V, а схема S содержит такие слова $V_2 = V_3R_1$ и $V_4 = R_1V_5$, что $V = (V_2,V_4)_{R_1} = V_3R_1V_5$ (см. рисунок).



Фрагмент схемы S *Фрагмент схемы S_2*

Слово $V = RV_1 = V_3R_1V_5$ имеет префиксы R и V_3. Рассмотрим три случая соотношения длин $|R|$ и $|V_3|$:

$$|R| = |V_3|, \quad |R| > |V_3| \quad \text{и} \quad |R| < |V_3|.$$

В случае $|R| = |V_3|$ имеем цепочку равенств

$$Z = (U,V)_R = (U_1R, V_3R_1V_5)_R = U_1RR_1V_5 = U \bullet V_4,$$

из которой следует, что слово Z может быть получено конкатенацией слов U и V_4. В случае $|R| > |V_3|$ возникает такое непустое R_2, что

$Z = (U, V_4)_{R_2}$. В обоих случаях при построении слова Z слово V_2 не используется. Поэтому его можно исключить из схемы S, что противоречит оптимальности схемы S. Следовательно, $|R| < |V_3|$ и $V_3 = RV_6$, $|V_6| \geqslant 1$. Теперь слова V_2 и V можно задать в виде

$$V_2 = V_3R_1 = RV_6R_1, \quad V = (V_2, V_4)_{R_1} = (RV_6R_1, R_1V_5)_{R_1} = RV_6R_1V_5$$

и получить новое представление слова Z:

$$Z = (U, V)_R = (U_1R, RV_6R_1V_5)_R = U_1RV_6R_1V_5. \tag{1}$$

Заменим в схеме S слово V на слово $(U, V_2)_R = U_1RV_6R_1$. В результате получим схему S_2, содержащую такие слова $(U, V_2)_R = U_1RV_6R_1$ и $V_4 = R_1V_5$, что

$$Z = ((U, V_2)_R, V_4)_{R_1} = U_1RV_6R_1V_5.$$

Полученное представление слова Z совпадает с (1). Схема S_2 оптимальна, поскольку число слов в ней равно числу слов в схеме S.

В схеме S_2 суффикс $V_4 = R_1V_5$ слова Z короче суффикса $V = RV_6R_1V_5$ слова Z в схеме S. Если суффикс V_4 слова Z не удовлетворяет утверждению леммы, то описанная выше процедура повторяется до тех пор, пока вновь полученный суффикс уменьшенной длины не удовлетворит утверждению леммы. Лемма доказана.

Теорема 1. *Для любого слова W выполняются неравенства*

$$L(W) \geqslant L_{sh}(W) \geqslant L^*(W) \geqslant L_i^{**}(W).$$

Доказательство. Так как операция композиции является обобщением операции конкатенации, то для всякого слова W очевидно неравенство $L(W) \geqslant L_{sh}(W)$. Неравенство $L^*(W) \geqslant L_i^{**}(W)$ содержится в предложении 2. В [3] получено неравенство $L(W) \geqslant L^*(W)$. Докажем справедливость неравенства $L_{sh}(W) \geqslant L^*(W)$.

Предположим, что теорема не верна. Из множества слов, для которых теорема не верна, выберем слово минимальной сложности. Пусть таким словом является Z и $L_{sh}(Z) < L^*(Z)$. Согласно лемме среди оптимальных схем композиции слова Z существует схема S, содержащая такие слова $U = U_1R$ и $V = RV_1$, что $Z = (U, V)_R = U_1RV_1$ (возможно, R пусто) и слово V является либо подсловом слова U, либо символом, отсутствующим в слове U. Пусть $Z = XY$, где Y — максимальный суффикс слова Z. Установим ряд неравенств.

Если подсхема S_1, $S_1 \subset S$, композиции слова U оптимальна, то $L_{sh}(Z) = L_{sh}(U) + 1$. В противном случае $L_{sh}(Z) > L_{sh}(U) + 1$. Поэтому

$$L_{sh}(Z) \geqslant L_{sh}(U) + 1. \tag{2}$$

Из минимальности выбранной по предположению схемы следует, что

$$L_{sh}(U) \geqslant L^*(U). \tag{3}$$

Если R пусто и V — символ, отсутствующий в слове U, то по определению слово V является максимальным суффиксом, $U = X$ и $L^*(U) = L^*(X)$. Если $V \sqsubseteq U$, то $V_1 \sqsubseteq Y$, $X \sqsubseteq U$ и согласно предложению 1

$$L^*(U) \geqslant L^*(X). \tag{4}$$

По определению суффиксного представления слова Z имеем $L^*(X) = L^*(Z) - 1$. Отсюда и из (2)–(4) следует, что

$$L_{sh}(Z) \geqslant L_{sh}(U) + 1 \geqslant L^*(U) + 1 \geqslant L^*(X) + 1 = L^*(Z).$$

Это противоречит сделанному предположению. Теорема 1 доказана.

Расширение возможностей схем композиции по сравнению со схемами конкатенации иногда позволяет получить простое решение. Например, при построени слов в однобуквенном алфавите $\Sigma = \{a\}$. В настояшее время окончательное решение этой задачи в классе схем конкатенации не получено. А для схем композиции слова a^t, где $2^{n-1} < t \leqslant 2^n$, $L_{sh}(a^t) = n$. Это следует из того, что за $n-1$ операцию слово длины более 2^{n-1} получить нельзя, а сложность схемы $a, a^{2^1}, a^{2^2}, \ldots, a^{2^{n-1}}, a^t$ равна n.

Пусть $2^{n-1} < t_1 < \ldots < t_m \leqslant 2^n$, $n \geqslant 1$. Для слова a^{t_1} имеем $L_{sh}(a^{t_1}) = n$. Для слов $a^{t_2}, \ldots, a^{t_m}$ справедливы равенства $a^{t_i} = (a^{t_{i-1}}, a^{t_{i-1}})_{R_i}$, $2 \leqslant i \leqslant m$. Следовательно, при любом натуральном n имеем $L_{sh}(a^{t_1}, \ldots, a^{t_m}) = n + m - 1$.

§ 2. Сложность схем композиции мономов

Пусть $\hat{U} = a_1^{u_1} \ldots a_q^{u_q}$, $\hat{V} = a_1^{v_1} \ldots a_q^{v_q}$ — произвольные мономы (здесь и далее мономы будем отмечать галочкой) и задан $\hat{R} = a_1^{r_1} \ldots a_q^{r_q}$, где $0 \leqslant r_i \leqslant \min(u_i, v_i), 1 \leqslant i \leqslant q$. Операция *композиции* мономов $\hat{U}$ и $\hat{V}$ относительно $\hat{R}$ определяется как $a_1^{u_1+v_1-r_1} \ldots a_q^{u_q+v_q-r_q}$ и обозначается через $(\hat{U}, \hat{V})_{\hat{R}}$. При пустом $\hat{R}$ получаем операцию коммутативной конкатенации $\hat{U} \bullet \hat{V}$.

Последовательность $a_1 \ldots, a_q, \hat{X}, \hat{Y}, \ldots, \hat{Z}$, которую обозначим через $\mathcal{S}$, назовем *схемой композиции* монома $\hat{Z}$, если для любого $\hat{W}$ из $\mathcal{S}$, начиная с $\hat{X}$, в $\mathcal{S}$ имеются такие $\hat{U}, \hat{V}$ (возможно, $\hat{U} = \hat{V}$), предшествующие $\hat{W}$, и существует такое $\hat{R}$, что $\hat{W} = (\hat{U}, \hat{V})_{\hat{R}}$. Под *сложностью* $L_{sh}(\mathcal{S})$ схемы $\mathcal{S}$ композиции $\hat{Z}$ понимается число мономов в последовательности $\hat{X}, \hat{Y}, \ldots, \hat{Z}$. Пусть $L_{sh}(\hat{Z}) = \min L_{sh}(\mathcal{S})$, где минимум берется по всем схемам композиции монома $\hat{Z}$. Величину $L_{sh}(\hat{Z})$ назовем *сложностью* монома $\hat{Z}$.

Схема $\mathcal{S}$ композиции монома $\hat{W} = a_1^{k_1} \ldots a_q^{k_q}$, $k_1, \ldots, k_q \geqslant 1$, назовем *$r$-ограниченной,* если в каждой операции $\hat{W}_i = (\hat{U}, \hat{V})_{\hat{R}_i}$, $\hat{W}_i, \hat{U}, \hat{V} \in \mathcal{S}$, используется моном $\hat{R}_i$, в котором степень каждой буквы не превосходит r. При $r = 0$ получаем схему коммутативной конкатенации.

Теорема 2. *Для любого монома $\hat{W} = a_1^{k_1} \ldots a_q^{k_q}$, $k_1, \ldots, k_q \geqslant 1$, порожденного 1-ограниченной схемой, справедливо равенство*

$$L_{sh}(\hat{W}) = \lceil \log_2 k \rceil + q - 1, \quad \textit{где} \quad k = \max\,(k_1, \ldots, k_q). \tag{5}$$

Доказательство. Н и ж н я я о ц е н к а. Для получения монома $a_1^{k_1}$ необходимо использовать не менее $\lceil \log_2 k_1 \rceil$ операций. Если моном $a_1^{k_1}$ уже имеется, то для получения монома $a_1^{k_1} a_2 \ldots a_q$ требуется использовать $q - 1$ конкатенаций. Следовательно, для получения монома $\hat{W} = a_1^{k_1} \ldots a_q^{k_q}$, где $k_1 \geqslant \ldots \geqslant k_q \geqslant 1$, требуется выполнить не менее $\lceil \log_2 k_1 \rceil + q - 1$ операций.

В е р х н я я о ц е н к а. Построим 1-ограниченную схему композиции монома $\hat{W}$. Пусть $k_1 \geqslant \ldots \geqslant k_q \geqslant 1$. Для каждого из сомножителей $a_1^{k_1}, \ldots, a_q^{k_q}$ построим суффиксное представление $a_i^{k_i} = X_{i,1} \ldots X_{i,\lceil \log_2 k_i \rceil + 1}$, $1 \leqslant i \leqslant q$. Все полученные максимальные суффиксы разместим в таблице $T_{\hat{W}}$, состоящей из q строк и $q + \lceil \log_2 k_1 \rceil$ столбцов. На каждой позиции таблицы $T_{\hat{W}}$ расположим не более одного суффикса. На позициях i-й строки последовательно поместим суффиксы $X_{i,1}, \ldots, X_{i,\lceil \log_2 k_i \rceil + 1}$. При заполнении таблицы $T_{\hat{W}}$ выполним два требования:

(i) в первых q столбцах таблицы $T_{\hat{W}}$ на главной диагонали разместим $X_{1,1}, \ldots, X_{q,1}$, оставив другие позиции первых q столбцов пустыми;

(ii) оставшиеся суффиксы разместим на позициях последних $\lceil \log_2 k_1 \rceil$ столбцов таблицы $T_{\hat{W}}$, причем в каждой строке все суффиксы разместим в конце строки.

Для каждого столбца таблицы $T_{\hat{W}}$ выполним конкатенацию всех слов данного столбца. В результате получим слова X_i, из которых построим слова $W_i = X_1 \ldots X_i$, $1 \leqslant i \leqslant q + \lceil \log_2 k_1 \rceil$. По построению имеем

$W_i = X_1 \dots X_i = W_{i-1}X_i$, $2 \leqslant i \leqslant q + \lceil \log_2 k_1 \rceil$, где кратности вхождения каждой буквы алфавита в слова W_{i-1} и X_i либо равны, либо в W_{i-1} имеется на одно вхождение больше чем в X_i. Поэтому $\hat{W}_i = \hat{W}_{i-1}a_{i-1}$, $2 \leqslant i \leqslant q$, $\hat{W}_i = (\hat{W}_{i-1}, \hat{W}_{i-1})_{\hat{R}_i}$, $q+1 \leqslant i \leqslant q + \lceil \log_2 k_1 \rceil$, где число вхождений любой буквы в $\hat{R}_i$ не превосходит единицы. Следовательно, существует схема $a_1, \dots, a_q, \hat{W}_2, \dots, \hat{W}_{q+\lceil \log_2 k_1 \rceil} = \hat{W}$. Теорема 2 доказана.

Пример. Проиллюстрируем описанный метод и укажем схему композиции монома $\hat{W} = a^{15}b^4c^3$ в алфавите $\{a, b, c\}$.

	a			a	a^2	a^4	a^7
		b				b	b^2
			c			c	c
$X_i:$	a	b	c	a	a^2	a^4bc	a^7b^2c
$W_i:$	a	ab	abc	a^2bc	a^4bc	$a^8b^2c^2$	$a^{15}b^4c^3$

Используя описанную выше процедуру, получаем $\hat{W}_2 = a \cdot b = ab$, $\hat{W}_3 = ab \cdot c = abc$, $\hat{W}_4 = (abc, abc)_{bc} = a^2bc$, $\hat{W}_5 = (a^2bc, a^2bc)_{bc} = a^4bc$, $\hat{W}_6 = a^4bc \cdot a^4bc = a^8b^2c^2$, $\hat{W}_7 = (a^8b^2c^2, a^8b^2c^2)_{ac} = a^{15}b^4c^3$ и схему $a, b, c, \hat{W}_2, \hat{W}_3, \hat{W}_4, \hat{W}_5, \hat{W}_6, \hat{W}_7$.

При доказательстве нижней оценки сложности ограничения теоремы 2 на моном $\hat{R}_i$ не используются. Поэтому равенство (5) справедливо и в общем случае, когда $r \geqslant 1$.

Выражаю благодарность А. А. Евдокимову за полезные обсуждения результатов и советы при написании статьи.

Литература

1. **Кочергин В. В.** О сложности вычислений в конечных абелевых группах // Докл. АН СССР. 1991. Т. 317, № 2. С. 291–294.
2. **Кочергин В. В.** О мультипликативной сложности двоичных слов с заданным числом единиц // Математические вопросы кибернетики. Вып. 8. М.: Наука. Физматлит, 1999. С. 63–76.
3. **Меркин Ю. В.** Нижняя оценка сложности для схем конкатенации слов // Дискрет. анализ и исслед. операций. 1996. Т. 3, № 1. С. 52–56.
4. **Меркин Ю. В.** Нижние оценки мультипликативной сложности символьных последовательностей, определяемых монотонными симметрическими булевыми функциями // Дискрет. анализ и исслед. операций. Сер. 1. 1999. Т. 6, № 3. С. 3–9.

5. **Мерекин Ю. В.** Нижние оценки сложности символьных последовательностей, определяемых симметрическими булевыми функциями // Дискрет. анализ и исслед. операций. Сер. 1. 2000. Т. 7, № 2. С. 54–64.

6. **Мерекин Ю. В.** Оценки мультипликативной сложности двоичных слов, определяемых поясковыми булевыми функциями // Дискрет. анализ и исслед. операций. Сер. 1. 2002. Т. 9, № 2. С. 36–47.

7. **Ширшов А. И.** Некоторые алгоритмические проблемы для алгебр Ли // Сиб. матем. журнал. 1962. Т. 3, № 2. С. 292–296.

Адрес автора:
Институт математики
им. С.Л.Соболева СО РАН,
пр. Академика Коптюга, 4,
630090 Новосибирск,
Россия.
E-mail: merekin@math.nsc.ru

Статья поступила
14 февраля 2003 г.
Переработанный вариант
7 октября 2003 г.