О длине минимальной алфавитной склейки для класса линейных регулярных языков

Дергач П.С., Раджабов Ж.И.

В кандидатской диссертации [1] была поставлена и решена задача о нахождении верхней оценки на минимальную длину слов из регулярного языка, склеивающихся (то есть имеющих совпадающий образ) при алфавитном кодировании (если такая склейка вообще существует). В данной статье исследуется задача о нахождении соответствующих нижних оценок на длину склейки для случая, когда регулярные языки имеют линейную функцию роста, а схема кодирования преобразует все буквы входного алфавита в один и тот же символ. Для такого кодирования образ слова однозначно определяется по его длине. Приводятся нижние оценки, совпадающие по порядку с верхними оценками из [1] для таких языков и такого кодирования. Кроме того, для этого подслучая приводится более точная верхняя оценка.

Ключевые слова: алфавитное кодирование, регулярный язык, склейка.

Введение

В работе [1] решается проблема проверки однозначности алфавитного декодирования в классе регулярных языков с некоторыми ограничениями на функцию роста. Вполне естественно, что для этого при условии существования склейки строятся верхние оценки на ее длину. Однако, вопрос о соответствующих нижних оценках тоже представляет отдельный научный интерес. Поскольку в общем случае решить эту задачу сложно, то для первого приближения было решено рассмотреть класс регулярных языков с линейной функцией роста со схемой кодирования, преобразующей все буквы входного алфавита в один и тот же символ.О решении похожих задач можно прочитать в статьях [2-8]. О других интересных аспектах исследований авторов и других ученых в смежных областях к тематике данной работы можно прочитать в [9-20].

Основные определения

Пусть $A = B = \{0, 1\}$. Пусть P_1, P_2 — непустые множества слов в алфавите A. Определим следующие операции над P_1 и P_2 :

1) *Объединение* множеств P_1 и P_2 (обозначение $P_1 \bigcup P_2$) есть множество всех слов вида α, где $\alpha \in P_1$ или $\alpha \in P_2$.

2) *Конкатенация* множеств P_1 и P_2 (обозначение $P_1 \cdot P_2$) есть множество всех слов вида $\alpha_1\alpha_2$, где $\alpha_1 \in P_1,\ \alpha_2 \in P_2$.

3) *Итерация* множества P_1 (обозначение $(P_1)^*$) есть множество всех слов вида $\alpha_1 \dots \alpha_k$, где $\alpha_1, \alpha_2, \dots, \alpha_k \in P_1,\ k \geqslant 0$. При $k = 0$ здесь имеется ввиду пустое слово λ.

Введем понятие регулярного языка в алфавите A. Называем множество $P,\ P \subseteq A^*$ *регулярным языком в алфавите* A, если его можно получить из пустого множества и одноэлементных однобуквенных множеств $\{a\},\ a \in A$ применением конечного числа конкатенаций, объединений и итераций. Более подробно, определение регулярных языков таково:

1) $\{a\}$, где a — произвольная буква алфавита A, — регулярные языки в алфавите A;

2) Если P_1, P_2 — регулярные языки в алфавите A, то и множества $P_1 \cup P_2$, $P_1 \cdot P_2$, $(P_1)^*$ — регулярные языки в алфавите A;

3) Регулярность произвольного языка в алфавите A устанавливается в соответствиями с пунктами (1)-(3) за конечное число шагов.

Множество регулярных языков в алфавите A обозначаем через $R(A)$.

Рассмотрим *схему алфавитного кодирования* $f : A \to B$, для которой $f(0) = f(1) = 0$. Далее эта схема кодирования доопределяется на произвольном языке $P \subset A^*$ следующим образом:

$$\tilde{f}(a_{i_1}a_{i_2} \dots a_{i_n}) = f(a_{i_1})f(a_{i_2}) \dots f(a_{i_n}) = 00 \dots 0.$$

Полученную функцию $\tilde{f} : A^* \to B^*$ называем *функцией алфавитного кодирования.*

Пусть $P \in R(A)$ и $\beta \in \tilde{f}(P)$. Тогда $\alpha \in P$ называется *расшифровкой* β *при алфавитном кодировании* $\tilde{f}$ *на регулярном языке* P или просто *расшифровкой* β, если $\tilde{f}(\alpha) = \beta$. Также говорим, что β — *код слова* α. Если для любых различных слов $\alpha_1, \alpha_2 \in P$ выполняется $\tilde{f}(\alpha_1) \neq \tilde{f}(\alpha_2)$,

то *декодирование однозначно на P по $\tilde{f}$*. Также говорим, что $\tilde{f}$ *биективно на P*. В противном случае говорим, что *в регулярном языке P есть склейка* (α_1, α_2). Под склейкой здесь понимается произвольная неупорядоченная пара различных слов языка P с одинаковым кодом. *Минимальной склейкой для языка P* называем склейку, доставляющую среди всех склеек языка P минимальное значение на максимум длин слов из склейки. А само это значение называем *размером минимальной склейки* и обозначаем его через $m(P)$. Для произвольной склейки ее *размером* также называем максимальную длину слов из этой склейки. Впрочем, очевидно, что для функции $\tilde{f}$ длины слов, образующих склейку, совпадают.

Пусть $\mathbf{E}$ — произвольное множество языков в алфавите A, каждый из которых имеет склейку. Через $m(\mathbf{E})$ обозначаем величину

$$m(\mathbf{E}) := \max_{P \in \mathbf{E}} m(P),$$

если, конечно, такой максимум существует.

Пусть $P \subseteq A^*$. Через $\mathbb{N}_0$ обозначаем множество $\mathbb{N} \cup \{0\}$. Для произвольного $n \in \mathbb{N}$ через $P_{\leqslant}(n)$ обозначаем множество слов из P, длина которых не превосходит n. Через $T_n(P)$ обозначаем мощность множества $P_{\leqslant}(n)$:

$$T_n(P) := |P_{\leqslant}(n)|.$$

Через T_P обозначаем функцию $T_P : \mathbb{N} \to \mathbb{N}_0$, где

$$T_P(n) := T_n(P)$$

для всех $n \in \mathbb{N}$. Называем T_P *функцией роста* для P. Говорим, что бесконечный язык P *имеет линейную функцию роста* и пишем $T_P \in Lin$, если функция T_P ограничена сверху каким-нибудь полиномом первой степени. Обозначаем класс бесконечных регулярных языков в алфавите A с линейной функцией роста через $LR(A)$. Из [1] известно, что всякий язык из класса $LR(A)$ представим в виде

$$\bigvee_{i=1}^{s} \alpha_i \beta_i^* \gamma_i, \tag{1}$$

где $\alpha_i, \beta_i, \gamma_i$ — слова в алфавите A и $\beta \neq \lambda$. Число s в этом представлении называется *его высотой*. Класс всех языков $P \in LR(A)$, представимых

выражением (1) с высотой s обозначим через $LR(A, s)$. *Сложностью представления (1)* называется число

$$L\left(\bigvee_{i=1}^{s} \alpha_i \beta_i^* \gamma_i\right) := \max_{i=1,\ldots,s} (|\tilde{f}(\alpha_i)| + |\tilde{f}(\beta_i)| + |\tilde{f}(\gamma_i)|).$$

Здесь под $|\alpha|$ имеется ввиду длина слова α, то есть количество букв в этом слове.

Сложностью языка $P \in LR(A, s)$ с линейной функцией роста называется минимальная сложность среди всех его представлений вида (1), имеющих высоту s. Для произвольного $k \in \mathbb{N}$ через $LR(A, s, k)$ обозначаем множество всех языков $P \in LR(A, s)$, имеющих сложность не выше k.

Утверждение 1. *Пусть* $k \in \mathbb{N}$, $k \geqslant 2$. *Тогда*

$$m(LR(A, 2, k)) \geqslant k(k-1).$$

Утверждение 2. *Пусть* $k > s > 2$. *Тогда*

$$m(LR(A, s, k)) \geqslant \left\rceil \frac{(k+2-s)(k+1-s)}{s-1} \right\lceil .$$

Утверждение 3. *Пусть* $k, s \in \mathbb{N}$, $k, s \geqslant 2$. *Тогда*

$$m(LR(A, s, k)) \leqslant 2k(k-1).$$

Доказательство утверждений

Лемма 1. *Пусть* $k \in \mathbb{N}$, $k \geqslant 2$, $P = \alpha_1 \beta_1^* \gamma_1 \vee \alpha_2 \beta_2^* \gamma_2$ *для некоторых* $\alpha_i, \beta_i, \gamma_i \in A^*$, $\beta_i \neq \lambda$, $|\alpha_i| + |\beta_i| + |\gamma_i| \leqslant k$ *и в* P *есть склейка. Тогда*

$$m(P) \leqslant 2k(k-1).$$

▷ **Доказательство леммы 1:**

Рассмотрим множество длин слов из языка P. Оно состоит из двух арифметических прогрессий l_1, l_2 с началом и шагом не выше k. Так как в P есть склейка, то эти прогрессии пересекаются. Отсюда, очевидно, следует, что они пересекаются по бесконечной арифметической прогрессии. Обозначим эту прогрессию через (a, b), где a — начало прогрессии, а b — ее шаг.

Покажем, что $a \leqslant k(k-1)$. Здесь возможны 2 случая. В первом из них оба шага прогрессий l_1, l_2 равны k и тогда $|\beta_1| = |\beta_2| = k$, а значит $\alpha_1 = \alpha_2 = \gamma_1 = \gamma_2 = \lambda$. Тогда прогрессии l_1, l_2 равны (k, k) и их пересечение (a, b) тоже равно (k, k). Очевидно, что $a = k \leqslant k(k-1)$. Во втором случае хотя бы один из шагов прогрессий l_1, l_2 меньше k и тогда утверждение следует из китайской теоремы об остатках, ведь НОК прыжков прогрессий l_1, l_2 в этом случае не превосходит $k(k-1)$.

Покажем, что $b \leqslant k(k-1)$. В первом случае (смотри выше) прогрессия (a, b) равна (k, k) и $b = k \leqslant k(k-1)$ — утверждение очевидно. Во втором случае утверждение, опять же, следует из китайской теоремы об остатках и того факта, что НОК прыжков прогрессий l_1, l_2 не превосходит $k(k-1)$.

Рассмотрим теперь два слова ρ_1, ρ_2 из языков $\alpha_1\beta_1^*\gamma_1$, $\alpha_2\beta_2^*\gamma_2$ соответственно, которые имеют длину a. Либо они образуют склейку (и в этом случае утверждение леммы очевидно), либо они совпадают. Если они, все-таки совпадают, то рассмотрим два других слова ρ_3, ρ_4 из языков $\alpha_1\beta_1^*\gamma_1$, $\alpha_2\beta_2^*\gamma_2$ соответственно, которые уже имеют длину $a + b$. Покажем от противного, что они образуют склейку. Для этого нам потребуется более внимательно посмотреть на структуру слов ρ_i. Пусть

$$\rho_1 = \alpha_1\beta_1^x\gamma_1,\ \rho_2 = \alpha_2\beta_1^y\gamma_2,\ \rho_3 = \alpha_1\beta_1^{x+z}\gamma_1,\ \rho_4 = \alpha_2\beta_2^{y+w}\gamma_2.$$

Введем ряд дополнительных обозначений

$$\gamma_1' := \alpha_1\beta_1^x,\ \gamma_2' := \alpha_2\beta_2^y,\ \delta_1 := \beta_1^z,\ \delta_2 := \beta_2^w.$$

Тогда получаем

$$\gamma_1'\gamma_1 = \rho_1 = \rho_2 = \gamma_2'\gamma_2,\ \gamma_1'\delta_1\gamma_1 = \rho_3 = \rho_4 = \gamma_2'\delta_2\gamma_2. \tag{2}$$

Без ограничения общности, $|\gamma_1| \leqslant |\gamma_2|$. Тогда

$$\gamma_2 = \gamma_3\gamma_1 \tag{3}$$

для некоторого (возможно, пустого) γ_3. Условия (2) с учетом (3) можно переписать в виде

$$\gamma_1' = \gamma_2'\gamma_3, \tag{4.1}$$

$$\gamma_1'\delta_1 = \gamma_2'\delta_2\gamma_3. \tag{4.2}$$

Подставив (4.1) в (4.2), получаем

$$\gamma_3\delta_1 = \delta_2\gamma_3. \tag{4.3}$$

Мы знаем, что у P есть какая-то склейка (ρ_5, ρ_6). Очевидно, что тогда она состоит из двух слов вида

$$\gamma_1'\delta_1^c\gamma_1 = \rho_5,\ \gamma_2'\delta_2^c\gamma_2 = \rho_6,$$

где $c \geqslant 2$. Но тогда из (3), (4.1-4.3) выводим

$$\rho_5 = \gamma_1'\delta_1^c\gamma_1 = \gamma_2'\gamma_3\delta_1^c\gamma_1 = \gamma_2'\gamma_3\delta_1\delta_1^{c-1}\gamma_1 = \gamma_2'\delta_2\gamma_3\delta_1^{c-1}\gamma_1 = \ldots =$$

$$= \gamma_2'\delta_2^{c-1}\gamma_3\delta_1\gamma_1 = \gamma_2'\delta_2^c\gamma_3\gamma_1 = \gamma_2'\delta_2^c\gamma_2 = \rho_6.$$

Полученное противоречие доказывает, что (ρ_3, ρ_4) — склейка. Осталось вспомнить, что размер этой склейки совпадает с длиной слов (ρ_3, ρ_4) и значит равен $a + b$. Но мы уже доказали, что $a, b \leqslant k(k-1)$. Поэтому $a + b \leqslant 2k(k-1)$. ■

Утверждение 1. *Пусть* $k \in \mathbb{N}$, $k \geqslant 2$. *Тогда*

$$m(LR(A, 2, k)) \geqslant k(k-1).$$

▷ **Доказательство утверждения 1:**

Для доказательства утверждения достаточно привести пример такого множества $P \in LR(A, 2, k)$, размер минимальной склейки в котором не меньше $k(k-1)$. Это верно, в частности для

$$P := (0^k)^* \cup (1^{k-1})^*,$$

так как $\text{НОК}(k, k-1) = k(k-1)$. ■

Утверждение 2. *Пусть* $k > s > 2$. *Тогда*

$$m(LR(A, s, k)) \geqslant \left]\frac{(k+2-s)(k+1-s)}{s-1}\right[.$$

▷ **Доказательство утверждения 2:**

Обозначим через $t(s,k)$ число $\left]\frac{(k+2-s)(k+1-s)}{s-1}\right[$. Для доказательства утверждения достаточно привести пример такого множества P из класса $LR(A,s,k)$, размер минимальной склейки в котором не меньше $t(s,k)$. Рассмотрим язык

$$P := (01^{k-1-s}0)^* \cup (01^{k-s}0)^* \cup \ldots (01^{k-3}0)^* \cup (01^{k-2}0)^*. \qquad (5)$$

Очевидно, что он принадлежит $LR(A,s,k)$. И у него есть склейки. Рассмотрим произвольную склейку (α_1,α_2) в языке (5). Тогда для некоторых $1 \leqslant i < j \leqslant s$ верно, что

$$\alpha_1 \in (01^{k+1-i}0)^*,\ \alpha_2 \in (01^{k+1-j}0)^*.$$

Из определения размера склейки следует, что размер склейки (α_1,α_2) равен длине слов α_1 и α_2. Но длина слова α_1 делится нацело на $k+1-i$, а длина слова α_2 делится нацело на $k+1-j$. Поэтому размер склейки делится на НОК$(k+1-i,k+1-j)$. Осталось заметить, что

$$\text{НОК}(k+1-i,k+1-j) = \frac{(k+1-i)(k+1-j)}{\text{НОД}(k+1-i,k+1-j)} \geqslant$$

$$\geqslant \frac{(k+2-s)(k+1-s)}{j-i} \geqslant \frac{(k+2-s)(k+1-s)}{s-1}.$$

А значит верно и что

$$\text{НОК}(k+1-i,k+1-j) \geqslant \left]\frac{(k+2-s)(k+1-s)}{s-1}\right[= t(s,k).$$

Поэтому размер минимальной склейки тоже не меньше $t(s,k)$. ■

Утверждение 3. *Пусть* $k,s \in \mathbb{N},\ k,s \geqslant 2$. *Тогда*

$$m(LR(A,s,k)) \leqslant 2k(k-1).$$

▷ **Доказательство утверждения 3:**

Для доказательства утверждения достаточно показать, что размер минимальной склейки (если она есть) для произвольных языков из

класса $LR(A, s, k)$ не превосходит $2k(k-1)$. Рассмотрим любой язык $P \in LR(A, s, k)$. Значит

$$P = \bigvee_{i=1}^{s} \alpha_i \beta_i^* \gamma_i,$$

где $|\alpha_i| + |\beta_i| + |\gamma_i| \leqslant k$ при $i = 1, \ldots, s$. Пусть (δ_1, δ_2) — минимальная склейка языка P. Без ограничения общности, можно считать, что

$$\delta_1 \in \alpha_1 \beta_1^* \gamma_1, \ \delta_2 \in \alpha_2 \beta_2^* \gamma_2.$$

Тогда (δ_1, δ_2) будет минимальной склейкой и в множестве

$$P_1 := \alpha_1 \beta_1^* \gamma_1 \bigvee \alpha_2 \beta_2^* \gamma_2.$$

Для доказательства утверждения осталось применить лемму 1. ■

Список литературы

[1] П. С. Дергач. *Алфавитное кодирование регулярных языков с полиномиальной функцией роста*. Кандидатская диссертация, Москва, 2016.

[2] П. С. Дергач, Э. С. Айрапетов. *О прогрессивном разбиении некоторых подмножеств натурального ряда*. Интеллектуальные системы, 2015. Т.19, вып. 3, М., Сс. 79-86.

[3] П. С. Дергач. *О каноническом регулярном представлении S-тонких языков*. Интеллектуальные системы, 2014. Т.18, вып. 1, М., Сс. 211-242.системы, 2014. Т.18, вып. 1, М., Сс. 211-242.

[4] П. С. Дергач. *О проблеме вложения допустимых классов*. Интеллектуальные системы, 2015. Т.19, вып. 2, М., Сс. 143-174.

[5] П. С. Дергач. *О двух размерностях спектров тонких языков*. Интеллектуальные системы, 2015. Т.19, вып. 3, М., Сс. 155-174.

[6] П. С. Дергач, Э. С. Айрапетов. *О прогрессивном разбиении последовательности натуральных чисел, имеющей пропуск длины 2*. Интеллектуальные системы, 2016. Т.20, вып. 2, М., Сс. 67-86.

[7] П. С. Дергач, Е. Д. Данилевская. *О покрытиях и разбиениях натуральных чисел, имеющих два последовательных пропуска длины 1.* Интеллектуальные системы, 2017. Т.21, вып. 1, М., Сс.192-237.

[8] П. С. Дергач. *О структуре вложения прогрессивных множеств сложности два.* Интеллектуальные системы, 2017. Т.21, вып. 2, М., Сс.117-162.

[9] Д. Е. Александров. *Эффективные методы реализации проверки содержания сетевых пакетов регулярными выражениями.* Интеллектуальные системы, 2014. Т.18, вып. 1, М., Сс. 37-60.

[10] Д. Н. Бабин. *Частотные регулярные языки.* Интеллектуальные системы, 2014. Т.18, вып. 1, М., Сс. 205-210.

[11] Д. Е. Александров. *Об оценках автоматной сложности распознавания классов регулярных языков.* Интеллектуальные системы, 2014. Т.18, вып. 4, М., Сс. 161-190.

[12] В. М. Дементьев. *О звездной высоте регулярного языка и циклической сложности минимального автомата.* Интеллектуальные системы, 2014. Т.18, вып. 4, М., Сс. 215-222.

[13] И. Е. Иванов. *О сохранении периодических последовательностей автоматами с магазинной памятью с однобуквенным магазином.* Интеллектуальные системы, 2015. Т.19, вып. 1, М., Сс. 145-160.

[14] А. А. Петюшко. *О контекстно-свободных биграмных языках.* Интеллектуальные системы, 2015. Т.19, вып. 2, М., Сс. 187-208.

[15] И. Е. Иванов. *Нижняя оценка на максимальную длину периода выходной последовательности автономного автомата с магазинной памятью.* Интеллектуальные системы, 2015. Т.19, вып. 3, М., Сс. 175-194.

[16] В. А. Орлов. *О конечных автоматах с максимальной степенью различимости состояний.* Интеллектуальные системы, 2016. Т.20, вып. 1, М., Сс. 213-222.

[17] П. С. Дергач. *О проблеме проверки однозначности алфавитного декодирования в классе регулярных языков с полиномиальной функцией роста.* Интеллектуальные системы, 2016. Т.20, вып. 2, М., Сс. 147-202.

[18] А. М. Миронов. *Основные понятия теории вероятностных автоматов*. Интеллектуальные системы, 2016. Т.20, вып. 2, М., Сс. 283-330.

[19] А. А. Петюшко, Д. Н. Бабин. *Классификация Хомского для матриц биграммных языков*. Интеллектуальные системы, 2016. Т.20, вып. 2, М., Сс. 331-336.

[20] С. Б. Родин. *О связи линейно реализуемых автоматов и автоматов с максимальной вариативностью относительно кодирования состояний*. Интеллектуальные системы, 2016. Т.20, вып. 2, М., Сс. 337-348.

Сведения об авторах

Дергач Петр Сергеевич, Dergach Pyotr Sergeevich

Младший научный сотрудник МГУ имени М. В. Ломоносова в городе Москве;

адрес: Россия, г. Москва, 125565, Ленинградское ш., 88-19;

тел. моб.: +79037189288;

e-mail: dergachpes@mail.ru.

Раджабов Жахонгир Ихтиер угли, Radjabov Jakhongir Ikhtiyor ogli

Студент факультета ПМиИ филиала МГУ имени М. В. Ломоносова в городе Ташкенте;

адрес: Узбекистан, г. Ташкент, 100000, ул. Лашкарбеги, 1;

тел. моб.: +998977335030;

e-mail: karkidon@icloud.com.

On the length of a minimal alphabetical bonding in linear regular languages

Dergach P.S., Radjabov J.I.

In the Ph.D. thesis [1] it has been found the upper bound on the minimal length of two words in regular language with similar image under alphabetic coding (if such pair of words exists at all). In this paper, we investigate the problem of finding corresponding lower bounds in subcase when regular languages have a linear growth function, and the coding scheme transforms all the letters of the input alphabet into the same symbol. For such encoding, the image of any word is uniquely determined by its length. Below we give lower bounds that coincide in order with the upper bounds from [1] for such languages and such coding. In addition, a more accurate upper estimate is given for this subcase.

Keywords: alphabetic coding, regular language, bonding.