Об оценках автоматной сложности распознавания класса регулярных языков

Д. Е. Александров

Рассмотрен метод решения проблемы «экспоненциального взрыва» числа состояний конечного автомата, распознающего множество регулярных языков, задаваемых объединением регулярных выражений вида $.*\mathtt{R}_1.*\mathtt{R}_2.*$, где $\mathtt{R}_1$ и $\mathtt{R}_2$ — произвольные регулярные выражения. Предложено расширение этого метода на случай объединения произвольного числа регулярных выражений данного вида. Приведены оценки на число состояний автомата при таком изменении в случае алфавита, состоящего из не менее чем трех символов. Показывается, что относительное уменьшение числа состояний может быть произвольным. Анализируется практическая эффективность предложенного метода применительно к регулярным выражениям системы Snort.

Ключевые слова: конечные автоматы, регулярные выражения, системы обнаружения вторжений.

Введение

Одну из ключевых ролей в сфере информационных технологий играют экспертные системы, выносящие определенный вердикт относительно поданных на вход слов путем проверки принадлежности слова заранее заданному регулярному множеству. В частности, важное место в области информационной безопасности занимают сетевые системы обнаружения и (или) предотвращения вторжений, такие как Snort [1], Bro [2], L7-filter [3] и аппаратные продукты фирмы Cisco [4]. Все они имеют базы сигнатур — наборы регулярных выражений, задающие регулярные языки, слова которых признаются вредоносными.

Традиционно для проверки принадлежности слова заданному регулярному языку строится детерминированный конечный автомат, принимающий данный язык. Однако с ростом числа выражений в наборе возрастает пространственная сложность (число состояний) конечного автомата для проверки принадлежности слов заданному регулярному множеству.

Существует два основных подхода к решению данной задачи. Первый подход — модификация традиционных детерминированного и недетерминированного конечных автоматов. Так, например, в работе С. Кумара [5], в основе которой лежит алгоритм Ахо — Корасик [6], предлагается использовать ДКА с сокращенным определенным образом числом состояний. В работе [7] предлагается одновременно использовать два специально сформированных автомата — ДКА и НКА, благодаря которым объем требуемой для автомата памяти меньше, чем в случае одного ДКА. В работах [8, 9] предлагается вводить специальные счетчики и битовые флаги, изменяемые в случае определенных переходов между состояниями, что также сокращает необходимый объем памяти.

Второй подход предполагает изменение исходного набора регулярных выражений таким образом, чтобы сократить сложность реализации в конечных автоматах за счет расширения определяемого выражениями регулярного языка. Однако методы, предлагаемые в работах по данной теме, как например в статье [10], подразумевают лишь «ручное» переписывание конкретных выражений. В работе [11] предложен метод модификации произвольного набора выражений, принадлежащих классу выражений вида $.*\mathtt{R}_1.*\mathtt{R}_2.*$ (в нотации PCRE [12]), где $\mathtt{R}_1$ и $\mathtt{R}_2$ — произвольные регулярные выражения, и даны оценки на относительный выигрыш в числе состояний детерминированного конечного автомата для случая двух выражений.

В настоящей работе предложены оценки числа состояний автоматов для случая набора из более чем двух выражений из класса $.*\mathtt{R}_1.*\mathtt{R}_2.*$ при применении алгоритма статьи [11] к одной паре выражений из набора. Вначале кратко изложен метод расширения регулярного языка, заданного двумя выражениями, и приведены оценки на число состояний при таком подходе, описанные в статье [11]. Далее даны оценки на число состояний автоматов для исходного и модифицированного набора из более чем двух выражений. Затем даны

результаты применения описанного алгоритма к реально используемым выражениям.

Основные понятия и результаты

В дополнение к традиционному определению регулярного выражения [13, 14] как задания набора операций объединения, конкатенации и «звезды Клини» над некоторым набором символов далее будут использоваться следующие операции PCRE-совместимых регулярных выражений: «.» — символ, обозначающий объединение всех символов алфавита; «[$\mathtt{a_1a_2}\ldots\mathtt{a_n}$]» — обозначение объединения символов $a_1, a_2 \ldots a_n$; «[$\mathtt{\hat{}a_1a_2}\ldots\mathtt{a_n}$]» — обозначение объединения всех символов алфавита кроме символов $a_1, a_2 \ldots a_n$. Под регулярным языком над заданным алфавитом Σ далее будет подразумеваться подмножество множества Σ^*, описываемое некоторым регулярным выражением. Заметим, что вышеперечисленные операции не расширяют класс регулярных языков, так как новые операции над регулярными выражениями выводятся из базового набора операций, а значит классы регулярных языков, определяемых традиционными и расширенными регулярными выражениями, совпадают. Следовательно будет верна теорема Клини [14] о том, что множество слов является регулярным языком тогда и только тогда, когда существует детерминированный конечный автомат, распознающий его. Здесь и далее под детерминированным конечным автоматом подразумевается инициальный детерминированный конечный автомат без выхода, однозначно задаваемый пятеркой объектов $\langle Q, \Sigma, q_0, \delta, A\rangle$, где Q — конечное множество состояний, Σ — входной алфавит, $q_0 \in Q$ — начальное состояние, $\delta\colon Q \times \Sigma \to Q$ — функция перехода, а $A \subseteq Q$ — множество финальных состояний.

Пусть R — регулярное выражение над алфавитом Σ, $L(R)$ — регулярный язык, определяемый выражением R, $V(L(R))$ — приведенный ДКА, принимающий язык $L(R)$, а $|V|$ — число состояний автомата V. Через $L^{pf}(R)$ обозначим такое наибольшее подмножество $L(R)$, что ни одно из слов из $L(R)$ не является нетривиальным префиксом для слов из $L^{pf}(R)$, где под нетривиальным префиксом подразумевается префикс, не совпадающий со всем словом. В случае, если $L(R)$ содержит пустую строку Λ, положим $L^{pf}(R) = \{\Lambda\}$.

В статье [11] была описана проблема экспоненциального роста числа состояний при объединении регулярных выражений вида $.*\mathtt{R}_1.*\mathtt{R}_2.*$, а также предложен способ нивелирования данной проблемы. Пусть $\mathtt{R}^1 = .*\mathtt{R}_1.*\mathtt{R}_2.*$ и $\mathtt{R}^2 = .*\mathtt{R}_3.*\mathtt{R}_4.*$ — два регулярных выражения. Для сокращения числа состояний конечного автомата, распознающего принадлежность слова регулярному множеству $L\left(\mathtt{R}^1\right) \cup L\left(\mathtt{R}^2\right)$, предлагается расширить регулярное множество $L\left(\mathtt{R}^1\right) \cup L\left(\mathtt{R}^2\right)$ до множества $L\left(.*\left(\mathtt{R}_1|\mathtt{R}_3\right).*\left(\mathtt{R}_2|\mathtt{R}_4\right).*\right)$ и построить ДКА, распознающий его. И хотя новый автомат может ошибочно принимать слова не из языка $L\left(\mathtt{R}^1\right) \cup L\left(\mathtt{R}^2\right)$, он будет иметь не больше, а в некоторых случая заметно меньше состояний, чем автомат, принимающий язык $L\left(\mathtt{R}^1\right) \cup L\left(\mathtt{R}^2\right)$. В статье [11] приведены оценки на число состояний исходных и новых автоматов для случая набора из двух выражений. В следующей части будут даны оценки на число состояний распознающих автоматов для случая произвольного числа выражений заданного вида, причем описанная модификация выражений будет применяться только для одной пары выражений из набора. Для доказательства утверждений настоящей статьи будут использоваться следующие утверждения статьи [11]:

Лемма 1. *ДКА $V\left(L\left(.*\mathtt{R}.*\right)\right)$, где $\mathtt{R}$ — произвольное регулярное выражение, имеет одно финальное состояние, которое также является поглощающим, и не содержит нефинальных поглощающих состояний.*

Следствие 1. *Для произвольного регулярного выражения $\mathtt{R}$ верно, что*

$$\left|V\left(L\left(.*\mathtt{R}.*\right)\right)\right| + 1 = \left|V\left(L^{pf}\left(.*\mathtt{R}\right)\right)\right|.$$

Лемма 2. *Пусть заданы 2 регулярных выражения $\mathtt{R}_1$ и $\mathtt{R}_2$, тогда*

$$\left|V\left(L\left(.*\mathtt{R}_1.*\mathtt{R}_2.*\right)\right)\right| = \left|V\left(L^{pf}\left(.*\mathtt{R}_1\right)\right)\right| + \left|V\left(L^{pf}\left(.*\mathtt{R}_2\right)\right)\right| - 3.$$

Лемма 3. *Пусть R' и R'' — такие регулярные языки, что автоматы $V\left(L\left(\mathtt{R}'\right)\right)$ и $V\left(L\left(\mathtt{R}''\right)\right)$ имеют по одному финальному состоянию, причем оба финальных состояния являются поглощающими. Тогда* $\left|V\left(L\left(\mathtt{R}'\right) \cup L\left(\mathtt{R}''\right)\right)\right| \leqslant \left(\left|V\left(L\left(\mathtt{R}'\right)\right)\right| - 1\right) \cdot \left(\left|V\left(L\left(\mathtt{R}''\right)\right)\right| - 1\right) + 1$

Теорема 1. *Пусть имеется два регулярных выражения* $.*\mathtt{R}_1.*\mathtt{R}_2.*$ *и* $.*\mathtt{R}_3.*\mathtt{R}_4.*$ *такие, что* $\left|V\left(L^{pf}\left(.*\mathtt{R}_\mathtt{i}\right)\right)\right| = n_i + 2$, $n_i \geqslant 1$. *Тогда:*

$$\begin{aligned} |V\left(L\left(.*\mathtt{R}_1.*\mathtt{R}_2.*\right)\right)| &= n_1 + n_2 + 1, \\ |V\left(L\left(.*\mathtt{R}_3.*\mathtt{R}_4.*\right)\right)| &= n_3 + n_4 + 1, \end{aligned} \tag{1}$$

$$|V\left(L\left(.*\mathtt{R}_1.*\mathtt{R}_2.*\right) \cup L\left(.*\mathtt{R}_3.*\mathtt{R}_4.*\right)\right)| \leqslant (n_1+n_2)\cdot(n_3+n_4)+1, \tag{2}$$

$$|V\left(L\left(.*(\mathtt{R}_1|\mathtt{R}_3).*(\mathtt{R}_2|\mathtt{R}_4).*\right)\right)| \leqslant n_1\cdot n_3 + n_2\cdot n_4 + 1. \tag{3}$$

Причем в случае $|\Sigma| \geqslant 5$ *оценки (2) и (3), вообще говоря, неулучшаемы.*

Если $|\Sigma| \geqslant 3$, *то для любого* $C \in (0;1] \cap \mathbb{Q}$ *найдутся такие выражения* $\mathtt{R}_1$, $\mathtt{R}_2$, $\mathtt{R}_3$ *и* $\mathtt{R}_4$, *что:*

$$C = \frac{|V\left(L\left(.*(\mathtt{R}_1|\mathtt{R}_3).*(\mathtt{R}_2|\mathtt{R}_4).*\right)\right)|}{|V\left(L\left(.*\mathtt{R}_1.*\mathtt{R}_2.*\right) \cup L\left(.*\mathtt{R}_3.*\mathtt{R}_4.*\right)\right)|}. \tag{4}$$

Ниже сформулированы утверждения, доказательства которых предложены в настоящей статье. Пусть задано $n > 2$ регулярных выражений вида $\mathtt{R}^\mathtt{i} = .*\mathtt{R}_{2\cdot\mathtt{i}-1}.*\mathtt{R}_{2\cdot\mathtt{i}}.*$. Для оценки возможного числа состояний детерминированных конечных автоматов, распознающих регулярные языки $\bigcup\limits_{i=1}^{n} L\left(\mathtt{R}^\mathtt{i}\right)$ и $L\left(.*(\mathtt{R}_1|\mathtt{R}_3).*(\mathtt{R}_2|\mathtt{R}_4).*\right) \cup \bigcup\limits_{i=3}^{n} L\left(\mathtt{R}^\mathtt{i}\right)$, введем две следующих функции Шеннона:

$$S\left(n,\mathfrak{N}\right) = \max_{|V(L(.*\mathtt{R}_\mathtt{i}.*))|=n_i+1} \left|V\left(\bigcup_{i=1}^{n} L\left(\mathtt{R}^\mathtt{i}\right)\right)\right|,$$

$$S_e\left(n,\mathfrak{N}\right) = \max_{|V(L(.*\mathtt{R}_\mathtt{i}.*))|=n_i+1} \left|V\left(L\left(.*(\mathtt{R}_1|\mathtt{R}_3).*(\mathtt{R}_2|\mathtt{R}_4).*\right) \cup \bigcup_{i=3}^{n} L\left(\mathtt{R}^\mathtt{i}\right)\right)\right|,$$

где $\mathfrak{N}$ — набор чисел $n_j \geqslant 1, j = \overline{1,2n}$.

Теорема 2.

$$S\left(n,\mathfrak{N}\right) \leqslant \prod_{i=1}^{n}\left(n_{2i-1}+n_{2i}\right)+1, \tag{5}$$

Пусть $|\Sigma| \geqslant 3$ *и набор чисел* $\mathfrak{N}$ *таков, что любое число* n_i *из него делится на* $b_1 = \left\lceil \log_{|\Sigma|-1} n \right\rceil$. *Тогда верно следующее неравенство:*

$$S\left(n,\mathfrak{N}\right) \geqslant \frac{1}{b_1^n}\prod_{i=1}^{n}\left(n_{2i-1}+n_{2i}\right)+\left(\sum_{i=1}^{b_1-1}\left(|\Sigma|-1\right)^i\right)\prod_{i=1}^{n}\left(\frac{n_{2i-1}}{b_1}+\frac{n_{2i}}{b_1}-1\right)+1. \tag{6}$$

Пусть $|\Sigma| \geqslant 4$ *и все числа из набора* $\mathfrak{N}$ *не меньше* $b_2 = \left\lceil \log_{|\Sigma|-2} n \right\rceil$. *Тогда верно следующее неравенство:*

$$S(n,\mathfrak{N}) \geqslant \prod_{i=1}^{n} \left(\left\lfloor \frac{n_{2i-1}}{b_2} \right\rfloor + \left\lfloor \frac{n_{2i}}{b_2} \right\rfloor \right) + \\ + \left(\sum_{i=1}^{b_2-1} (|\Sigma|-2)^i \right) \prod_{i=1}^{n} \left(\left\lfloor \frac{n_{2i-1}}{b_2} \right\rfloor + \left\lfloor \frac{n_{2i}}{b_2} \right\rfloor - 1 \right) + 1. \quad (7)$$

Теорема 3. *Пусть* $n_j \geqslant 1$, $j = \overline{1,2n}$, $n \geqslant 2$. *Тогда верно неравенство:*

$$S_e(n,\mathfrak{N}) \leqslant (n_1 \cdot n_3 + n_2 \cdot n_4) \cdot \prod_{i=3}^{n} (n_{2i-1} + n_{2i}) + 1. \quad (8)$$

Пусть $|\Sigma| \geqslant 3$ *и набор чисел* $\mathfrak{N}$ *таков, что любое число* n_i *из него делится на* $b_1 = \left\lceil \log_{|\Sigma|-1} n \right\rceil$. *Тогда верно следующее неравенство:*

$$S_e(n,\mathfrak{N}) \geqslant \left(\frac{n_1}{b_1} \cdot \frac{n_3}{b_1} + \frac{n_2}{b_1} \cdot \frac{n_4}{b_1} \right) \cdot \prod_{i=3}^{n} \left(\frac{n_{2i-1}}{b_1} + \frac{n_{2i}}{b_1} \right) + \\ + \left(\sum_{i=1}^{b_1-1} (|\Sigma|-1)^i \right) \cdot \left(\frac{n_1}{b_1} \cdot \frac{n_3}{b_1} + \left(\frac{n_2}{b_1} - 1 \right) \cdot \left(\frac{n_4}{b_1} - 1 \right) \right) \cdot \\ \cdot \prod_{i=3}^{n} \left(\frac{n_{2i-1}}{b_1} + \frac{n_{2i}}{b_1} - 1 \right) + 1, \quad (9)$$

Пусть $|\Sigma| \geqslant 4$ *и все числа из набора* $\mathfrak{N}$ *не меньше* $b_2 = \left\lceil \log_{|\Sigma|-2} n \right\rceil$. *Тогда верно следующее неравенство:*

$$S_e(n,\mathfrak{N}) \geqslant \left(\left\lfloor \frac{n_1}{b_2} \right\rfloor \cdot \left\lfloor \frac{n_3}{b_2} \right\rfloor + \left\lfloor \frac{n_2}{b_2} \right\rfloor \cdot \left\lfloor \frac{n_4}{b_2} \right\rfloor \right) \cdot \prod_{i=3}^{n} \left(\left\lfloor \frac{n_{2i-1}}{b_2} \right\rfloor + \left\lfloor \frac{n_{2i}}{b_2} \right\rfloor \right) + \\ + \left(\sum_{i=1}^{b_2-1} (|\Sigma|-2)^i \right) \cdot \left(\left\lfloor \frac{n_1}{b_2} \right\rfloor \cdot \left\lfloor \frac{n_3}{b_2} \right\rfloor + \left(\left\lfloor \frac{n_2}{b_2} \right\rfloor - 1 \right) \cdot \left(\left\lfloor \frac{n_4}{b_2} \right\rfloor - 1 \right) \right) \cdot \\ \cdot \prod_{i=3}^{n} \left(\left\lfloor \frac{n_{2i-1}}{b_2} \right\rfloor + \left\lfloor \frac{n_{2i}}{b_2} \right\rfloor - 1 \right) + 1. \quad (10)$$

Теорема 4. *Если* $|\Sigma| \geqslant 3$, *то для любого* $C \in (0;1] \cap \mathbb{Q}$ *и* $n \geqslant 2$ *найдутся такие выражения* $\mathtt{R^i} = .*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*$, $i = \overline{1,n}$, *что:*

$$C = \frac{\left|V\left(L\left(.*(\mathtt{R_1}|\mathtt{R_3}).*(\mathtt{R_2}|\mathtt{R_4}).*\right) \cup \bigcup\limits_{i=3}^{n} L\left(\mathtt{R^i}\right)\right)\right|}{\left|V\left(\bigcup\limits_{i=1}^{n} L\left(\mathtt{R^i}\right)\right)\right|}. \tag{11}$$

Вспомогательные утверждения

Лемма 4. *Пусть* $\mathtt{R_1}$ *и* $\mathtt{R_2}$ *— такие регулярные выражения над алфавитом* Σ, *что* $L(\mathtt{R_1}) \subseteq \Sigma_1^* \setminus \{\Lambda\}$ *и* $L(\mathtt{R_2}) \subseteq \Sigma_2^* \setminus \{\Lambda\}$, *где* Σ_1 *и* Σ_2 *— непересекающиеся непустые подмножества множества* Σ. *Тогда верно равенство:*

$$|V\left(L\left(.*(\mathtt{R_1}|\mathtt{R_2}).*\right)\right)| = |V\left(L\left(.*\mathtt{R_1}.*\right)\right)| + |V\left(L\left(.*\mathtt{R_2}.*\right)\right)| - 2.$$

Доказательство. Рассмотрим приведенные ДКА $V' = \left\langle Q', \Sigma, q_0', \delta', A' = \left\{q_f'\right\}\right\rangle$ и $V'' = \left\langle Q'', \Sigma, q_0'', \delta'', A'' = \left\{q_f''\right\}\right\rangle$, распознающие языки $L(\mathtt{R_1})$ и $L(\mathtt{R_2})$ соответственно. По лемме 1 данные автоматы имеют вид, изображенный на рисунке 1.

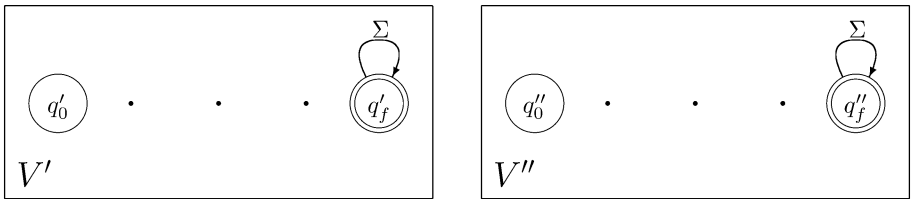

Рис. 1. Общий вид автоматов V' и V''.

Объединим эти два автомата в новый автомат V следующим образом: «сольем» состояния q_f' и q_f'' в финальное поглощающее состояние q_f. Объединим состояния q_0' и q_0'' в новое начальное состояние q_0, сохранив все переходы из q_0' в отличные от q_0' состояния и переходы из состояния q_0'' в отличные от q_0'' состояния. Для доказательства корректности такого объединения достаточно доказать, что

$$\forall q' \in Q' \setminus \left\{q_f'\right\}, a \in \Sigma \setminus \Sigma_1 \colon \delta'\left(q', a\right) = q_0' \tag{12}$$

$$\forall q'' \in Q'' \setminus \left\{q''_f\right\}, a \in \Sigma \setminus \Sigma_2 \colon \delta''\left(q'', a\right) = q''_0. \tag{13}$$

Действительно, тогда переходы из q'_0 в отличные от q'_0 состояния не могут осуществляться по символам из $\Sigma \setminus \Sigma_1$, а переходы из состояния q''_0 в состояния отличные от него не могут осуществляться по символам из $\Sigma \setminus \Sigma_2$. Поэтому переходы из q'_0, которые необходимо сохранить, имеют метки из множества Σ_1, а сохраняемые переходы из q''_0 — метки из Σ_2. По условию $\Sigma_1 \cap \Sigma_2 = \varnothing$, а значит объединение корректно.

Докажем утверждение (12) от противного. Пусть найдутся такие $q' \in Q' \setminus \left\{q'_f\right\}$ и $a \in \Sigma \setminus \Sigma_1$, что $q'_a = \delta'\left(q', a\right) \neq q'_0$. В силу того, что слова из языка $L\left(\mathtt{R_1}\right)$ не содержат символ a и q' отлично от q'_f, состояние q'_a также отлично от финального. По предположению состояния q'_a и q'_0 различны, а так как они являются состояниями приведенного автомата, они отличимы. То есть существует различающее их непустое слово $\alpha \in \Sigma^*$, переводящее одно из этих состояний в финальное состояние q'_f, а другое — в некоторое нефинальное. Пусть α переводит q'_a в финальное состояние. Обозначим через α_0 слово, переводящее состояние q'_0 в состояние q'. Тогда, по предположению, слово $\alpha_0 a \alpha$ должно переводить q'_0 в финальное состояние q'_f, то есть $\alpha_0 a \alpha \in L\left(.*\mathtt{R_1}.*\right) = \Sigma^* L\left(\mathtt{R_1}\right)\Sigma^*$. В силу того, что $a \in \Sigma \setminus \Sigma_1$, а $L\left(\mathtt{R_1}\right) \subseteq \Sigma_1^*$, либо слово α_0, либо слово α содержит некоторое слово из $L\left(\mathtt{R_1}\right)$. Заметим, что слово α_0 не может содержать слово из $L\left(\mathtt{R_1}\right)$, так как в противном случае α_0 переводит q'_0 в финальное состояние, которое отлично от q'. Следовательно α содержит слово из $L\left(\mathtt{R_1}\right)$, то есть принадлежит языку $L\left(.*\mathtt{R_1}.*\right)$. Но тогда слово α переводит состояния q'_0 и q'_a в финальное состояние q'_f. Противоречие. Пусть α переводит q'_0 в финальное состояние, а q'_a — нет. Если α_0 — слово, переводящее q'_0 в q'_a, то слово $\alpha_0 a \alpha$, по предположению, переводит q'_0 в нефинальное состояние. Однако $\alpha_0 a \alpha \in \Sigma^*\left(\Sigma \setminus \Sigma_1\right) L\left(.*\mathtt{R_1}.*\right) = \Sigma^*\left(\Sigma \setminus \Sigma_1\right)\Sigma^* L\left(\mathtt{R_1}\right)\Sigma^* \subseteq \Sigma^* L\left(\mathtt{R_1}\right)\Sigma^* = L\left(.*\mathtt{R_1}.*\right)$, а значит $\alpha_0 a \alpha$ переводит состояние q'_0 в финальное состояние. Но тогда, так как $\alpha_0 a$ переводит q'_0 в q'_a, то q'_a переводится словом α в финальное состояние. Противоречие. Значит утверждение (12) верно. А так как условия на выражения $\mathtt{R_1}$ и $\mathtt{R_2}$ схожи, то доказательство утверждения (13) аналогично.

Вернемся к объединению автоматов. Заменим в автомате V переходы из состояний $q' \in Q' \setminus \left\{q'_0, q'_f\right\}$ по символам $a \in \Sigma_2$ на переходы

в состояния $\delta''(q_0'', a)$, а переходы из состояний $q'' \in Q'' \setminus \left\{q_0'', q_f''\right\}$ по символам $a \in \Sigma_1$ на переходы в состояния $\delta'(q_0', a)$ соответственно.

Докажем, что автомат V принимает регулярный язык $L(.*(\mathtt{R}_1|\mathtt{R}_2).*)$. Заметим, что по построению автомат V неотличим от автомата V' на множестве слов $(\Sigma \setminus \Sigma_2)^*$, так как при слиянии были сохранены все переходы автомата V' по символам из $\Sigma \setminus \Sigma_2$. Аналогично V неотличим от V'' на множестве $(\Sigma \setminus \Sigma_1)^*$. Возьмем произвольное слово α из регулярного языка $L(.*(\mathtt{R}_1|\mathtt{R}_2).*)$. Без ограничения общности будем считать, что $\alpha \in L(.*\mathtt{R}_1.*)$. Пусть $\alpha \in (\Sigma \setminus \Sigma_2)^* L(\mathtt{R}_1)\Sigma^*$, то есть $\alpha = \alpha'\alpha''$, где $\alpha' \in L(.*\mathtt{R}_1) \cap (\Sigma \setminus \Sigma_2)^*$ и $\alpha'' \in \Sigma^*$. Как было замечено выше, V и V' неотличимы на множестве $(\Sigma \setminus \Sigma_2)^*$, а значит слово α' переводит q_0 в финальное поглощающее состояние q_f. Следовательно автомат V принимает слово $\alpha = \alpha'\alpha''$. Пусть теперь $\alpha \in \Sigma^* L(\mathtt{R}_1)\Sigma^* \setminus (\Sigma \setminus \Sigma_2)^* L(\mathtt{R}_1)\Sigma^*$. Тогда слово α можно представить в виде $\alpha = \alpha'\alpha''$, где $\alpha' \in \Sigma^*\Sigma_2$, $\alpha'' \in (\Sigma \setminus \Sigma_2)^* L(\mathtt{R}_1)\Sigma^*$. Если α' переводит q_0 в финальное состояние, то $\alpha = \alpha'\alpha''$ принимается автоматом V. Пусть α' переводит q_0 в некоторое нефинальное состояние q'. Заметим, что переходы по символам из Σ_2 ведут только в состояния $\{q_0, q_f\} \cup Q'' \setminus \left\{q_0'', q_f''\right\}$, поэтому нефинальное состояние $q' \in \{q_0\} \cup Q'' \setminus \left\{q_0'', q_f''\right\}$. По построению и из утверждения (13) следует, что при любом $a \in \Sigma \setminus \Sigma_2$ переходы из состояний из $\{q_0\} \cup Q'' \setminus \left\{q_0'', q_f''\right\}$ ведут в $\delta'(q_0', a)$. Поэтому любое слово из $(\Sigma \setminus \Sigma_2)\Sigma^*$ переводит состояния из $Q'' \setminus \left\{q_0'', q_f''\right\}$ туда же, куда и состояние q_0. По доказанному выше, слово $\alpha'' \in (\Sigma \setminus \Sigma_2)^* L(\mathtt{R}_1)\Sigma^*$ переводит состояние q_0 автомата V в финальное состояние, а значит α'' также переводит состояние q' в финальное состояние q_f. Следовательно $\alpha = \alpha'\alpha''$ принимается автоматом V.

Пусть теперь слово α принимается автоматом V. Так как единственное финальное состояние является поглощающим, в слове α можно выделить префикс α' минимальной длины, принимаемый автоматом V. Докажем, что $\alpha' \in L(.*(\mathtt{R}_1|\mathtt{R}_2).*)$. В силу того, что α' — минимальный префикс слова α, переводящий q_0 в q_f, путь из состояния q_0, соответствующий слову α', завершается в финальном состоянии q_f, но больше не проходит через него. Но тогда по построению и из утверждений (12) и (13) следует, что последний

символ α', переводящий в финальное состояние q_f из нефинального, принадлежит множеству $\Sigma_1 \cup \Sigma_2$. Без ограничения общности будем считать, что последний символ слова α' принадлежит Σ_1. Пусть $\alpha' \in (\Sigma \setminus \Sigma_2)^* \Sigma_1$. Тогда, так как автоматы V и V' неотличимы на множестве $(\Sigma \setminus \Sigma_2)^*$, слово α' будет приниматься автоматом V', а значит $\alpha' \in L(.*\mathtt{R_1}.*)$, но тогда $\alpha \in L(.*\mathtt{R_1}.*)\Sigma^* = L(.*R_1.*)$. Пусть $\alpha' \in \Sigma^*\Sigma_1 \setminus (\Sigma \setminus \Sigma_2)^* \Sigma_1$. Тогда α' можно представить в виде $\beta'\beta''$, где $\beta' \in \Sigma^*\Sigma_2$, $\beta'' \in (\Sigma \setminus \Sigma_2)^* \Sigma_1$. Аналогично предыдущим рассуждениям слово β' переводит состояние q_0 автомата V в некоторое состояние $q' \in \{q_0\} \cup Q'' \setminus \left\{q_0'', q_f''\right\}$, которое переводится словом β'' туда же, куда и состояние q_0. Поэтому β'' переводит q_0 в финальное состояние. В силу неотличимости автоматов V и V' на множестве $(\Sigma \setminus \Sigma_2)^*$, слово β'' также принимается автоматом V', то есть $\beta'' \in L(.*\mathtt{R_1}.*)$. Тогда $\alpha \in \Sigma^*\Sigma_2 L(.*\mathtt{R_1}.*)\Sigma^* \subseteq L(.*\mathtt{R_1}.*)$. Таким образом автомат V принимает язык $L(.*(\mathtt{R_1}|\mathtt{R_2}).*)$.

Докажем неприводимость автомата V. Состояние q_f отличимо от остальных состояний как единственное финальное состояние. Из утверждения (12) и неприводимости автомата V' следует, что исходные состояния $Q' \setminus \left\{q_f'\right\}$ автомата V' отличимы словами из Σ_1^*. В силу того, что при объединении автоматов V' и V'' в автомат V сохранились все переходы автомата V' по символам из Σ_1, состояния $\{q_0\} \cup Q' \setminus \left\{q_0', q_f'\right\}$ также попарно отличимы словами из Σ_1. Аналогично состояния $\{q_0\} \cup Q'' \setminus \left\{q_0'', q_f''\right\}$ попарно отличимы словами из Σ_2^*. Докажем отличимость состояний из $Q' \setminus \left\{q_0', q_f'\right\}$ и $Q'' \setminus \left\{q_0'', q_f''\right\}$. Как уже было замечено, состояния из $Q'' \setminus \left\{q_0'', q_f''\right\}$ переводятся непустыми словами из $(\Sigma \setminus \Sigma_2)^*$ туда же, куда и состояние q_0. То есть состояния из $Q'' \setminus \left\{q_0'', q_f''\right\}$ неотличимы от от состояния q_0 на множестве слов $(\Sigma \setminus \Sigma_2)^* \setminus \{\Lambda\}$. Поэтому слова из $\Sigma_1^* \subseteq (\Sigma \setminus \Sigma_2)^*$, отличающие состояния из $Q' \setminus \left\{q_0', q_f'\right\}$ от состояния q_0, также отличают их от состояний из $Q'' \setminus \left\{q_0'', q_f''\right\}$. Таким образом автомат V неприводим.

По построению число состояний приведенного автомата V, принимающего регулярный язык $L(.*(\mathtt{R_1}|\mathtt{R_2}).*)$, равняется

$$\left|V'\right| + \left|V''\right| - 2 = |V(L(.*\mathtt{R_1}.*))| + |V(L(.*\mathtt{R_2}.*))| - 2.$$

Лемма 5. *Пусть $L \subseteq \Sigma^*$ — регулярный язык над алфавитом Σ, $V_\Sigma(L)$ — автомат приведенного вида над алфавитом Σ, принимающий язык L, $V_{\Sigma'}(L \cap \Sigma'^*)$ — автомат приведенного вида над алфавитом $\Sigma' \subseteq \Sigma$, принимающий язык $L \cap \Sigma'^*$. Тогда верно следующее неравенство:*

$$|V_\Sigma(L)| \geqslant \left|V_{\Sigma'}\left(L \cap \Sigma'^*\right)\right|.$$

Доказательство. Обозначим $V_\Sigma = V_\Sigma(L)$, $V_{\Sigma'} = V_{\Sigma'}(L \cap \Sigma'^*)$.

Пусть V — автомат над алфавитом Σ', полученный из автомата V_Σ путем удаления переходов по символам из множества $\Sigma \setminus \Sigma'$. По построению переходы по символам из алфавита Σ' автоматов V и V_Σ совпадают, а значит они неотличимы на множестве слов из Σ'^*. Следовательно автомат V принимает регулярный язык $L \cap \Sigma'^*$, а значит $|V| \geqslant |V_{\Sigma'}|$, так как автомат $V_{\Sigma'}$ по условию является приведенным. При этом по построению $|V| = |V_\Sigma|$. Поэтому $|V_\Sigma| \geqslant |V_{\Sigma'}|$.

Лемма 6. *Пусть $L \subseteq \Sigma^*$ — некоторый регулярный язык над алфавитом Σ, $V_\Sigma(\Sigma^* L \Sigma^*)$ — автомат приведенного вида над алфавитом Σ, принимающий язык $\Sigma^* L \Sigma^*$, $V_{\Sigma'}(\Sigma'^* L \Sigma'^*)$ — автомат приведенного вида над алфавитом $\Sigma' \supseteq \Sigma$, принимающий язык $\Sigma'^* L \Sigma'^*$. Тогда верно следующее равенство:*

$$V_\Sigma(\Sigma^* L \Sigma^*) = V_{\Sigma'}\left(\Sigma'^* L \Sigma'^*\right).$$

Доказательство. Обозначим $V_\Sigma = V_\Sigma(\Sigma^* L \Sigma^*)$, $V_{\Sigma'} = V_{\Sigma'}(\Sigma'^* L \Sigma'^*)$.

Если язык L содержит пустое слово, то оба автомата V_Σ и $V_{\Sigma'}$ имеют всего по одному состоянию, которые также являются финальными поглощающими состояниями.

Докажем лемму для случая $\{\Lambda\} \notin L$. Пусть V — автомат над алфавитом Σ', полученный из автомата V_Σ путем добавления переходов из нефинальных состояний в начальное состояние по символам из $\Sigma' \setminus \Sigma$ и переходов из финального состояния в себя по символам из $\Sigma' \setminus \Sigma$. Докажем, что автомат V принимает язык $\Sigma'^* L \Sigma'^*$. Пусть $\alpha \in \Sigma'^* L \Sigma'^*$. Тогда α представимо в виде $\alpha = \alpha_1\alpha_2\alpha_3$, где $\alpha_1 \in (\Sigma'^*(\Sigma' \setminus \Sigma)) \cup \{\Lambda\}$, $\alpha_2 \in \Sigma^* L \Sigma^*$, $\alpha_3 \in ((\Sigma' \setminus \Sigma)\Sigma'^*) \cup \{\Lambda\}$. Предположим, что слово α_1 не переводит автомат V в финальное состояние (в противном случае автомат V, очевидно, принимает слово α), но тогда либо α_1 пусто и, следовательно, оставляет автомат V в начальном состоянии, либо α_1 непусто и его последний символ принадлежит множеству $\Sigma' \setminus \Sigma$, тогда по построению α_1 переводит автомат V

в начальное состояние. По построению переходы по символам из алфавита Σ автоматов V и V_Σ совпадают, а значит они неотличимы на множестве слов из Σ^*. Поэтому, так как автомат V_Σ принимает слово $\alpha_2 \in \Sigma^* L \Sigma^* \subset \Sigma^*$, слово α_2 также переводит и начальное состояние автомата V в финальное состояние. В силу того, что α_1 переводит автомат V в начальное состояние, то слово $\alpha_1\alpha_2$ переводит автомат V в финальное состояние. Так как финальное состояние является поглощающим, то $\alpha_1\alpha_2\alpha_3$ также переводит автомат V в финальное состояние. Следовательно автомат V принимает произвольное слово $\alpha \in \Sigma'^* L \Sigma'^*$. Пусть α — слово, принимаемое автоматом V. Выделим минимальный префикс α' слова α, принимаемый автоматом V. По построению в финальное состояние могут вести только переходы из финального состояния по символам из Σ' и переходы из нефинальных состояний по символам из Σ. Следовательно последний символ слова α' может принадлежать только алфавиту Σ. Тогда α' представимо в виде $\alpha' = \alpha'_1\alpha'_2$, где $\alpha'_1 \in (\Sigma'^* (\Sigma' \setminus \Sigma)) \cup \{\Lambda\}$, $\alpha'_2 \in \Sigma^*\Sigma$. По построению, так как никакой из префиксов слова α'_1 не переводит автомат V в финальное состояние, слово α'_1 переводит начальное состояние автомата V в себя, если непусто. Поэтому, так как $\alpha'_1\alpha'_2$ переводит автомат V в финальное состояние, слово α'_2 также переводит начальное состояние автомата V в финальное. В силу того, что автоматы V и V_Σ неотличимы на множестве Σ^*, а $\alpha'_2 \in \Sigma^*\Sigma \subset \Sigma^*$, слово α'_2 также переводит начальное состояние автомата V_Σ в финальное. Таким образом $\alpha'_2 \in \Sigma^* L \Sigma^*$. Поэтому слово α принадлежит языку $((\Sigma'^* (\Sigma' \setminus \Sigma)) \cup \{\Lambda\}) \Sigma^* L \Sigma^* \Sigma'^* \subseteq \Sigma'^* L \Sigma'^*$. Так как автоматы V и $V_{\Sigma'}$ принимают один и тот же язык, а автомат $V_{\Sigma'}$ приведен, верно неравенство $|V| \geqslant |V_{\Sigma'}|$, причем по построению $|V| = |V_\Sigma|$.

По лемме 5, в силу того, что $\Sigma'^* L \Sigma'^* \cap \Sigma^* = \Sigma^* L \Sigma^*$, получаем, что $|V_{\Sigma'}| \geqslant |V_\Sigma|$. Таким образом $|V_\Sigma| = |V| \geqslant |V_{\Sigma'}| \geqslant |V_\Sigma|$. Следовательно $|V_\Sigma| = |V_{\Sigma'}|$.

Лемма 7. *Пусть заданы $n \geqslant 2$ регулярных выражений $\mathtt{R_i}$, тогда*

$$|V\left(L\left(.*\mathtt{R_1}.*\ \ldots\ .*\mathtt{R_n}.*\right)\right)| = \sum_{i=1}^{n} \left|V\left(L^{pf}\left(.*\mathtt{R_i}\right)\right)\right| - 2n + 1.$$

Доказательство. Докажем утверждение индукцией по числу выражений. По лемме 2 равенство верно для двух выражений. Пусть равенство верно для $n-1$ выражения. Докажем равенство для

n выражений. Обозначим $\mathtt{R}_1^{\mathtt{n-1}} = \mathtt{R}_1.* \ \ldots \ .*\mathtt{R}_{\mathtt{n-1}}$. По предположению $\left|V\left(L\left(.*\mathtt{R}_1^{\mathtt{n-1}}.*\right)\right)\right| = \sum\limits_{i=1}^{n-1}\left|V\left(L^{pf}\left(.*\mathtt{R_i}\right)\right)\right| - 2\cdot(n-1)+1$. Применим лемму 2 к выражениям $\mathtt{R}_1^{\mathtt{n-1}}$ и $\mathtt{R_n}$. Тогда $\left|V\left(L\left(.*\mathtt{R}_1^{\mathtt{n-1}}.*\mathtt{R_n}.*\right)\right)\right| = \left|V\left(L^{pf}\left(.*\mathtt{R}_1^{\mathtt{n-1}}\right)\right)\right| + \left|V\left(L^{pf}\left(.*\mathtt{R_n}\right)\right)\right| - 4 + 1$. По следствию 1 $\left|V\left(L^{pf}\left(.*\mathtt{R}_1^{\mathtt{n-1}}\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R}_1^{\mathtt{n-1}}.*\right)\right)\right| + 1$, а значит $\left|V\left(L\left(.*\mathtt{R}_1^{\mathtt{n-1}}.*\mathtt{R_n}.*\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R}_1^{\mathtt{n-1}}.*\right)\right)\right| + 1 + \left|V\left(L^{pf}\left(.*\mathtt{R_n}\right)\right)\right| - 3 = \sum\limits_{i=1}^{n-1}\left|V\left(L^{pf}\left(.*\mathtt{R_i}\right)\right)\right| - 2\cdot(n-1)+2+\left|V\left(L^{pf}\left(.*\mathtt{R_n}\right)\right)\right| - 3 = \sum\limits_{i=1}^{n}\left|V\left(L^{pf}\left(.*\mathtt{R_i}\right)\right)\right| - 2n + 1$.

Доказательство теоремы 2

Докажем утверждение (5) индукцией по числу объединяемых выражений n. По следствию 1 $\left|V\left(L^{pf}\left(.*\mathtt{R_j}\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R_j}.*\right)\right)\right| + 1 = n_j + 2$. Тогда по лемме 2 $\left|L\left(\mathtt{R^i}\right)\right| = \left|L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)\right| = n_{2i-1} + n_{2i} + 1$, причем по лемме 1 автомат $V\left(L\left(\mathtt{R^i}\right)\right)$ при любом i удовлетворяет условию леммы 3. Пусть $n = 2$. По лемме 3 $\left|V\left(L\left(\mathtt{R^1}\right)\cup L\left(\mathtt{R^2}\right)\right)\right| \leqslant \left(\left|V\left(L\left(\mathtt{R^1}\right)\right)\right| - 1\right)\cdot\left(\left|V\left(L\left(\mathtt{R^2}\right)\right)\right| - 1\right) + 1 = (n_1+n_2)\cdot(n_3+n_4)+1$, причем по доказательству леммы 3 автомат $V\left(L\left(\mathtt{R^1}\right)\cup L\left(\mathtt{R^2}\right)\right)$ также удовлетворяет условию леммы 3. Пусть при $n = k$ автомат $V\left(\bigcup\limits_{i=1}^{k} L\left(\mathtt{R^i}\right)\right)$ имеет не более $\prod\limits_{i=1}^{k}(n_{2i-1}+n_{2i})+1$ состояний и удовлетворяет условию леммы 3. Рассмотрим случай $n = k+1$. Заметим, что $\bigcup\limits_{i=1}^{k+1} L\left(\mathtt{R^i}\right) = \bigcup\limits_{i=1}^{k} L\left(\mathtt{R^i}\right)\cup L\left(\mathtt{R^{k+1}}\right)$. По предположению автоматы $V\left(\bigcup\limits_{i=1}^{k} L\left(\mathtt{R^i}\right)\right)$ и $V\left(L\left(\mathtt{R^{k+1}}\right)\right)$ удовлетворяют условию леммы 3, следовательно $\left|V\left(\bigcup\limits_{i=1}^{k} L\left(\mathtt{R^i}\right)\cup L\left(\mathtt{R^{k+1}}\right)\right)\right| \leqslant \left(\left|V\left(\bigcup\limits_{i=1}^{k} L\left(\mathtt{R^i}\right)\right)\right| - 1\right)\cdot\left(\left|V\left(L\left(\mathtt{R^{k+1}}\right)\right)\right| - 1\right)+1 = \prod\limits_{i=1}^{k+1}(n_{2i-1}+n_{2i})+1$, причем автомат $V\left(\bigcup\limits_{i=1}^{k+1} L\left(\mathtt{R^i}\right)\right)$ также удовлетворяет условию леммы 3. Таким образом при всех $n \geqslant 2$ число состояний автомата $V\left(\bigcup\limits_{i=1}^{n} L\left(\mathtt{R^i}\right)\right)$ не превосходит $\prod\limits_{i=1}^{n}(n_{2i-1}+n_{2i})+1$.

Докажем утверждение (6). Пусть r — некоторый символ алфавита Σ, а $a_1, \ldots, a_{|\Sigma|-1}$ — различные символы из $\Sigma \setminus \{r\}$. Упорядочим множество слов $(\Sigma \setminus \{r\})^{b_1}$ в лексикографическом порядке справаналево (пример упорядочения слов из множества $\{0,1\}^3$: $000 < 100 < 010 < 110 < 001 \ldots$), считая, что $a_i < a_j$, если $i < j$. Множество $(\Sigma \setminus \{r\})^{b_1}$ содержит ровно $(|\Sigma|-1)^{b_1}$ слов, подставив значение b_1 получим $(|\Sigma|-1)^{\lceil \log_{|\Sigma|-1} n \rceil} \geqslant (|\Sigma|-1)^{\log_{|\Sigma|-1} n} = n$. Следовательно из этого множества можно выбрать первые n слов и обозначить их как $\alpha_1, \ldots, \alpha_n$. Для каждого $1 \leqslant i \leqslant n$ в качестве выражений $\mathtt{R_{2i-1}}$ и $\mathtt{R_{2i}}$ возьмем выражения вида $\underbrace{\alpha_i.* \ \ldots \ .*\alpha_i}_{\frac{n_{2i-1}}{b_1}}$ и $\underbrace{\alpha_i.* \ \ldots \ .*\alpha_i}_{\frac{n_{2i}}{b_1}}$ соответственно.

Докажем, что верно равенство $\left|L^{pf}\left(.*\mathtt{R_j}\right)\right| = n_j + 2$ для $j = \overline{1,2n}$. По лемме 7 и следствию 1 $\left|L^{pf}\left(.*\mathtt{R_j}\right)\right| = \left|V\left(L^{pf}\left(.*\alpha_\mathtt{i}.* \ \ldots \ .*\alpha_\mathtt{i}\right)\right)\right| = \left|V\left(L\left(.*\alpha_\mathtt{i}.* \ \ldots \ .*\alpha_\mathtt{i}.*\right)\right)\right| + 1 = \dfrac{n_j}{b_1} \cdot \left|V\left(L^{pf}\left(.*\alpha_j\right)\right)\right| - 2\dfrac{n_j}{b_1} + 2$. Пусть $\alpha_j = a_{i_1} \ldots a_{i_{b_1}}$, где $a_{i_1}, \ldots, a_{i_{b_1}}$ — символы алфавита $\Sigma \setminus \{r\}$ Тогда, очевидно, автомат приведенного вида, принимающий язык $L\left(.*\alpha_\mathtt{j}.*\right)$, имеет вид, изображенный на рисунке 2 (некоторые состояния и переходы не показаны для лучшей читаемости). Отметим, что по аналогии с доказательством леммы 4, все переходы из нефинальных состояний по символам $\{r\}$ ведут в начальное состояние данного автомата.

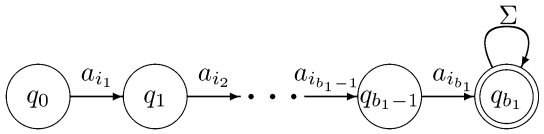

Рис. 2. Общий вид автомата $V\left(L\left(.*\alpha_j.*\right)\right)$.

Таким образом, по следствию 1 $\left|V\left(L^{pf}\left(.*\alpha_j\right)\right)\right| = \left|V\left(L\left(.*\alpha_\mathtt{j}.*\right)\right)\right| + 1 = b_1 + 2$. Поэтому $\left|L^{pf}\left(.*\mathtt{R_j}\right)\right| = \dfrac{n_j}{b_1} \cdot (b_1+2) - 2\dfrac{n_j}{b_1} + 2 = n_j + 2$. Докажем теперь, что автомат $V\left(\bigcup\limits_{i=1}^{n} L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)\right)$ имеет не менее чем $\dfrac{1}{b_1^n}\prod\limits_{i=1}^{n}\left(n_{2i-1}+n_{2i}\right) + \left(\sum\limits_{i=1}^{b_1-1}\left(|\Sigma|-1\right)^i\right) \cdot \prod\limits_{i=1}^{n}\left(\dfrac{n_{2i-1}}{b_1} + \dfrac{n_{2i}}{b_1} - 1\right) + 1$ состояний. Пусть $V_i = \left\langle Q_i, \Sigma, q_0^i, \delta_i, A_i = \left\{q_f^i\right\}\right\rangle$ — автомат, принимаю-

щий язык $L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)$, где $i=\overline{1,n}$. Пронумеруем состояния автоматов в соответствии с их глубиной, считая, что начальные состояния имеют номер 0. Тогда по построению из доказательств лемм 2 и 7 автомат V_i имеет вид, изображенный на рисунке 3, где группы переходов исходных автоматов, реализующие переходы по словам α_i из начальных состояний исходных автоматов в финальные, для большей наглядности изображены пунктирными переходами с метками α_i.

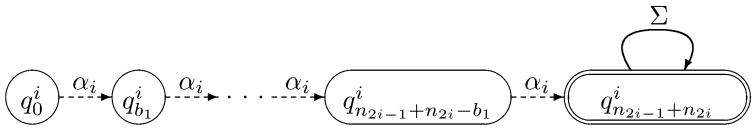

Рис. 3. Общий вид автомата V_i.

Построим автомат V с $\prod\limits_{i=1}^{n}|Q_i|$ состояниями, где каждому состоянию ставится в соответствие набор состояний $\left(q_{j_1}^1,\ldots,q_{j_n}^n\right)$, где $q_{j_i}^i\in Q_i$. Для произвольного символа $a\in\Sigma$ зададим значение функции переходов как $\delta\left(\left(q_{j_1}^1,\ldots,q_{j_n}^n\right),a\right)=\left(\delta_1\left(q_{j_1}^1,a\right),\ldots,\delta_n\left(q_{j_n}^n,a\right)\right)$. В качестве начального состояния возьмем состояние $\left(q_0^1,\ldots,q_0^n\right)$, а в качестве финальных — все состояния вида $\left(q_{j_1}^1,\ldots,q_{j_n}^n\right)$, где найдется такое i, что $q_{j_i}^i=q_f^i$. Заметим, что по построению все финальные состояния являются поглощающими, а значит и неотличимыми. Такой автомат очевидно будет принимать язык $\bigcup\limits_{i=1}^{n}L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)$, как автомат, реализующий по сути параллельную работу n автоматов V_i.

Назовем состояния вида $\left(q_{j_1}^1,\ldots,q_{j_n}^n\right)$, где $b\mid j_i,\ q_{j_i}^i\neq q_f^i=q_{n_{2i-1}+n_{2i}}^i$ при $i=\overline{1,n}$, «узловыми» состояниями. Докажем, что узловые состояния, которые по определению не являются финальными, достижимы и отличимы. Доказательство проведем индукцией по сумме индексов состояний $\sum\limits_{i=1}^{n}j_i$. Состояние $\left(q_0^1,\ldots,q_0^n\right)$ является начальным, а следовательно достижимо. Пусть достижимы узловые состояния, у которых $\sum\limits_{i=1}^{n}j_i\leqslant k\cdot b_1$. Возьмем такое произвольное узловое состояние $\left(q_{j_1}^1,\ldots,q_{j_n}^n\right)$, что $\sum\limits_{i=1}^{n}j_i=(k+1)\cdot b_1$. Из определения узло-

вого состояния следует, что найдется такое i, что $j_i \geqslant b_1$. По предположению состояние $\left(q^1_{j_1}, \ldots, q^i_{j_i - b_1}, \ldots, q^n_{j_n}\right)$ достижимо, так как сумма индексов состояний составляет $k \cdot b_1$. Слово $\alpha_i r$, очевидно, переводит состояние $q^i_{j_i - b_1}$ автомата V_i в состояние $q^i_{j_i}$, а остальные состояния $q^1_{j_1}, \ldots, q^n_{j_n}$ соответствующих автоматов переводит в себя. Следовательно, состояние $\left(q^1_{j_1}, \ldots, q^n_{j_n}\right)$ достижимо. Поэтому все узловые состояния достижимы. Пусть $\left(q^1_{j'_1}, \ldots, q^n_{j'_n}\right)$ и $\left(q^1_{j''_1}, \ldots, q^n_{j''_n}\right)$ — два различных узловых состояния. Без ограничения общности будем считать, что $j'_1 > j''_1$. Тогда, очевидно, слово $\underbrace{\alpha_1 r \ldots \alpha_1 r}_{\frac{n_1+n_2-j'_1}{b_1}}$ переводит состояние $\left(q^1_{j'_1}, \ldots, q^n_{j'_n}\right)$ в состояние $\left(q^1_{n_1+n_2}, q^2_{j'_2} \ldots, q^n_{j'_n}\right)$, а состояние $\left(q^1_{j''_1}, \ldots, q^n_{j''_n}\right)$ — в состояние $\left(q^1_{j''_1+n_1+n_2-j'_1}, q^2_{j''_2} \ldots, q^n_{j''_n}\right)$. По построению первое состояние является финальным, второе — нет, так как $n_1 + n_2 + j''_1 - j'_1 < n_1 + n_2$. Таким образом автомат содержит не менее $\dfrac{1}{b_1^n} \prod\limits_{i=1}^{n} (n_{2i-1} + n_{2i})$ узловых состояний.

Пусть $b_1 = 1$. Тогда, так как помимо узловых состояний, автомат содержит финальное состояние, утверждение (6) верно.

Пусть $b_1 > 1$. Рассмотрим отдельно узловые состояния, которые не могут быть переведены в финальное состояние словом длины b_1, то есть состояния $\left(q^1_{j_1}, \ldots, q^n_{j_n}\right)$, где $b_1 \mid j_i$, $j_i < n_{2i-1} + n_{2i} - b_1$ при $i = \overline{1, n}$. Докажем, что словами длины не более $b_1 - 1$ из каждого такого состояния достижимо не менее $\sum_{i=1}^{b_1-1} (|\Sigma| - 1)^i$ состояний отличимых друг от друга и от узловых состояний. Пусть $q = \left(q^1_{j_1}, \ldots, q^n_{j_n}\right)$ — некоторое такое узловое состояние. Далее через $q_\alpha = \left(q^1_\alpha, \ldots, q^n_\alpha\right)$ будем обозначать состояние, в которое по слову α переходит состояние q. На рисунке 4

изображены переходы из состояния q по непустым словам длины не более $b_1 - 1$ из множества $(\Sigma \setminus \{r\})^* = \left(\bigcup\limits_{i=1}^{|\Sigma|-1} \{a_i\}\right)^*$. Отметим, что другие узловые состояния не достижимы словами длины меньше b_1. Докажем, что все изображенные состояния q_α отличимы. В соответствии с выбором b_1 и слов α_i, для любого слова $\alpha \in (\Sigma \setminus \{r\})^{b_1-1}$ найдется такое $1 \leqslant i \leqslant n$, что $\alpha a_1 = \alpha_i$. Возьмем $q_{\alpha'}$ и $q_{\alpha''}$ — два из

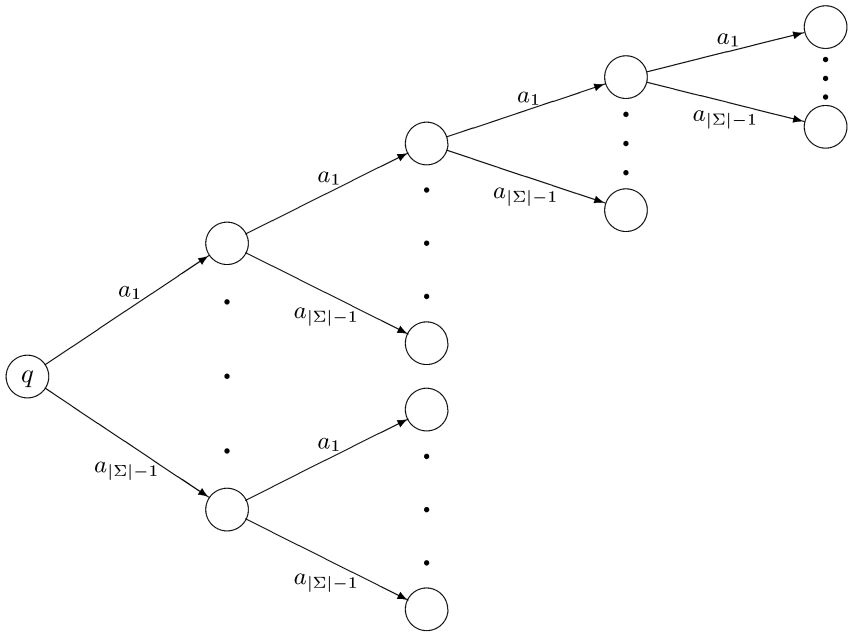

Рис. 4. Переходы из состояния q.

рассматриваемых состояния, причем $\alpha' \neq \alpha''$. Пусть длины α' и α'' совпадают. Как было отмечено выше, найдется такое i', что $\alpha_{i'} = \alpha'\beta'$, где β' — непустое слово из $(\Sigma \setminus \{r\})^*$ длины $b_1 - |\alpha'|$. Тогда, так как слово $\alpha_{i'}r$, очевидно, переводит состояние $q = \left(q^1_{j_1}, \ldots, q^{i'}_{j_i+b_1}, \ldots, q^n_{j_n}\right)$ в узловое состояние $\left(q^1_{j_1}, \ldots, q^{i'}_{j_{i'}+b_1}, \ldots, q^n_{j_n}\right)$, слово $\beta'r$ также переводит $q_{\alpha'}$ в это узловое состояние. Пусть найдется такое i'', что $\alpha_{i''} = \alpha''\beta'$. В силу того, что $\alpha' \neq \alpha''$, слово $\alpha_{i'}$ также не совпадает с $\alpha_{i''}$. Аналогично предыдущим рассуждениям слово $\alpha_{i''}r$ переводит q в $\left(q^1_{j_1}, \ldots, q^{i''}_{j_{i''}+b_1}, \ldots, q^n_{j_n}\right)$, а $\beta'r$ переводит состояние $q_{\alpha''}$ в него же. Таким образом $\beta'r$ переводит состояния $q_{\alpha'}$ и $q_{\alpha''}$ в отличимые узловые состояния. Следовательно они также отличимы. Если не найдется такое i'', что $\alpha_{i''} = \alpha''\beta'$, то $\beta'r$ переводит $q_{\alpha''}$ в состояние q. Узловые состояния $\left(q^1_{j_1}, \ldots, q^{i'}_{j_{i'}+b_1}, \ldots, q^n_{j_n}\right)$ и q отличимы, а значит состояния $q_{\alpha'}$ и $q_{\alpha''}$, переводимые в них словом $\beta'r$, также отличимы.

Пусть теперь длины α' и α'' отличаются. Без ограничения общности будем считать, что длина α' больше. Найдется такое i, что $\alpha_i = \alpha'\beta'$, где β' — непустое слово из $(\Sigma \setminus \{r\})^*$ длины $b_1 - |\alpha'|$. Аналогично предыдущим рассуждениям $\beta' r$ переводит $q_{\alpha'}$ в узловое состояние $\left(q^1_{j_1}, \ldots, q^i_{j_i+b_1}, \ldots, q^n_{j_n}\right)$, отличное от q. При этом слово $\alpha''\beta' r$ переводит состояние q в себя, так как по предположению длина $\alpha''\beta'$ меньше b_1. Узловые состояния, в которые переводятся словом $\beta' r$ состояния $q_{\alpha'}$ и $q_{\alpha''}$, отличимы, а значит $q_{\alpha'}$ и $q_{\alpha''}$ также отличимы.

Заметим, что рассматриваемые состояния, достижимые словами длины не более $b_1 - 1$ из различных узловых состояний, отличимы, так как переход по символу r переводит их в соответствующие узловые состояния, которые по доказанному ранее отличимы.

Таким образом число состояний автомата приведенного вида, распознающего язык $\bigcup\limits_{i=1}^{n} L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)$ с выражениями R_j, заданными выше, не меньше, чем

$$\prod_{i=1}^{n}\left(\frac{n_{2i-1}}{b_1} + \frac{n_{2i}}{b_1}\right) + \left(\sum_{i=1}^{b_1-1}(|\Sigma|-1)^i\right)\prod_{i=1}^{n}\left(\frac{n_{2i-1}}{b_1} + \frac{n_{2i}}{b_1} - 1\right) + 1,$$

где первое слагаемое — число узловых состояний, второе — число неузловых состояний, соответствующих $\prod\limits_{i=1}^{n}\left(\frac{n_{2i-1}}{b_1} + \frac{n_{2i}}{b_1} - 1\right)$ узловым состояниям, а последнее слагаемое — одно финальное состояние.

Докажем оценку (7). Пусть a — некоторый символ алфавита Σ. Введем следующие обозначения: $\Sigma' = \Sigma \setminus \{a\}$, $b'_1 = \left\lceil \log_{|\Sigma'|-1} n \right\rceil = \left\lceil \log_{|\Sigma|-2} n \right\rceil = b_2$ и $n'_i = \left\lfloor \frac{n_i}{b'_1} \right\rfloor \cdot b'_1$. Тогда, согласно доказательству утверждения (6), найдутся такие выражения $\mathtt{R'_j}$, $j = \overline{1,2n}$ над алфавитом Σ', что $\left|V\left(L^{pf}\left(.*\mathtt{R'_i}\right)\right)\right| = n'_i + 2$ и $\left|V\left(\bigcup\limits_{i=1}^{n} L\left(.*\mathtt{R'_{2i-1}}.*\mathtt{R'_{2i}}.*\right)\right)\right| \geqslant \prod\limits_{i=1}^{n}\left(\frac{n'_{2i-1}}{b'_1} + \frac{n'_{2i}}{b'_1}\right) + \left(\sum\limits_{i=1}^{b'_1-1}(|\Sigma'|-1)^i\right) \cdot \prod\limits_{i=1}^{n}\left(\frac{n'_{2i-1}}{b'_1} + \frac{n'_{2i}}{b'_1} - 1\right) + 1$. В качестве выражений $\mathtt{R_i}$ возьмем выражения $\mathtt{R'_i}$, если $n_i = n'_i$, и $(\mathtt{R'_i}|a\{n_i - n'_i + 1\})$ в противном случае. Отметим, что выражение вида $\mathtt{R'}.*\mathtt{R''}$ над алфавитом $\Sigma \setminus \{a\}$ совпадает с выражением $\mathtt{R'}[\mathtt{c_1} \ldots \mathtt{c_{|\Sigma|-1}}]\mathtt{R''}$, где $c_1, \ldots, c_{|\Sigma|-1}$ — различные символы алфавита Σ, отличные от символа a. Таким обра-

зом, выражения вида $\mathtt{R}'.*\mathtt{R}''.*\ldots.*\mathtt{R}^{(\mathtt{k})}$ над алфавитом $\Sigma\backslash\{a\}$ при переходе к алфавиту Σ примут вид $\mathtt{R}'[\hat{}\,a]*\mathtt{R}''[\hat{}\,a]*\ldots[\hat{}\,a]*\mathtt{R}^{(\mathtt{k})}$. По лемме 6 число состояний автомата $V\left(L\left(.*\mathtt{R}'_{\mathtt{i}}.*\right)\right)$ над алфавитом Σ совпадает с числом состояний автомата $V\left(L\left(.*\mathtt{R}'_{\mathtt{i}}.*\right)\right)$ над алфавитом Σ'. Применив следствие 1, получаем, что $\left|V\left(L^{pf}\left(.*\mathtt{R}'_{\mathtt{i}}\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R}'_{\mathtt{i}}.*\right)\right)\right|+1 = n'_i+2$, где $V\left(L^{pf}\left(.*\mathtt{R}'_{\mathtt{i}}\right)\right)$ — автомат над алфавитом Σ. При $n_i = n'_i$ очевидно, что $\left|V\left(L^{pf}\left(.*\mathtt{R}_{\mathtt{i}}\right)\right)\right| = \left|V\left(L^{pf}\left(.*\mathtt{R}'_{\mathtt{i}}\right)\right)\right| = n'_i+2$. В случае $n_i \neq n'_i$ по лемме 4, в силу того, что выражение $\mathtt{R}'_{\mathtt{i}}$ определено над алфавитом $\Sigma\setminus\{a\}$, верно равенство $\left|V\left(L\left(.*(\mathtt{R}'_{\mathtt{i}}|a\{n_i-n'_i+1\}).*\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R}'_{\mathtt{i}}.*\right)\right)\right| + \left|V\left(L\left(.*a\{n_i-n'_i+1\}.*\right)\right)\right| - 2$. Очевидно, что $\left|V\left(L\left(.*a\{n_i-n'_i+1\}.*\right)\right)\right| = n_i - n'_i + 2$. Поэтому, применив следствие 1, получаем: $\left|V\left(L^{pf}\left(.*\mathtt{R}_{\mathtt{i}}\right)\right)\right| = \left|V\left(L\left(.*(\mathtt{R}'_{\mathtt{i}}|a\{n_i-n'_i+1\}).*\right)\right)\right| + 1 = \left|V\left(L^{pf}\left(.*\mathtt{R}'_{\mathtt{i}}\right)\right)\right| - 1 + n_i - n'_i + 2 - 2 + 1 = n'_i + n_i - n'_i + 2 = n_i + 2$.

Заметим, что регулярный язык, принимаемый автоматом $V' = V\left(\bigcup\limits_{i=1}^{n} L\left(.*\mathtt{R}'_{\mathtt{2i-1}}.*\mathtt{R}'_{\mathtt{2i}}.*\right)\right)$ над алфавитом Σ', совпадает с языком $\left(\bigcup\limits_{i=1}^{n} L\left(.*\mathtt{R}_{\mathtt{2i-1}}.*\mathtt{R}_{\mathtt{2i}}.*\right)\right)\cap(\Sigma')^*$. Поэтому по лемме 5

$$\left|V\left(\bigcup_{i=1}^{n} L\left(.*\mathtt{R}_{\mathtt{2i-1}}.*\mathtt{R}_{\mathtt{2i}}.*\right)\right)\right| \geqslant \left|V'\right| \geqslant \prod_{i=1}^{n}\left(\frac{n'_{2i-1}}{b'_1}+\frac{n'_{2i}}{b'_1}\right) + $$

$$+\sum_{i=1}^{b'_1-1}\left(\left|\Sigma'\right|-1\right)^i\prod_{i=1}^{n}\left(\frac{n'_{2i-1}}{b'_1}+\frac{n'_{2i}}{b'_1}-1\right)+1 = \prod_{i=1}^{n}\left(\left\lfloor\frac{n_{2i-1}}{b_2}\right\rfloor+\left\lfloor\frac{n_{2i}}{b_2}\right\rfloor\right)+$$

$$+\sum_{i=1}^{b_2-1}\left(|\Sigma|-2\right)^i\prod_{i=1}^{n}\left(\left\lfloor\frac{n_{2i-1}}{b_2}\right\rfloor+\left\lfloor\frac{n_{2i}}{b_2}\right\rfloor-1\right)+1$$

Доказательство теоремы 3

Докажем утверждение (8). Как и в случае теоремы 2, верно равенство $\left|V\left(L^{pf}\left(.*\mathtt{R}_{\mathtt{j}}\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R}_{\mathtt{j}}.*\right)\right)\right|+1 = n_j+2$. По теореме 1 $\left|V\left(L\left(.*(\mathtt{R}_1|\mathtt{R}_3).*(\mathtt{R}_2|\mathtt{R}_4).*\right)\right)\right| \leqslant n_1\cdot n_3 + n_2\cdot n_4 + 1$, причем по лемме 1 данный автомат удовлетворяет условию леммы 2. По теореме 2 $\left|V\left(\bigcup\limits_{i=3}^{n} L\left(\mathtt{R}^{\mathtt{i}}\right)\right)\right| \leqslant \prod\limits_{i=3}^{n}\left(n_{2i-1}+n_{2i}\right)+1$, причем данный автомат также удовлетворяет условию леммы 2. Поэтому по лемме 2 верно следующее неравенство:

$$\left|V\left(L\left(.*(\mathtt{R_1}|\mathtt{R_3}).*(\mathtt{R_2}|\mathtt{R_4}).*\right)\cup\bigcup_{i=3}^{n}L\left(\mathtt{R^i}\right)\right)\right|\leqslant$$

$$\leqslant(n_1\cdot n_3+n_2\cdot n_4)\cdot\prod_{i=3}^{n}(n_{2i-1}+n_{2i})+1.$$

Доказательство утверждения (9) схоже с соответствующим доказательством теоремы 2

Пусть r — некоторый символ алфавита Σ, а $a_1,\ldots,a_{|\Sigma|-1}$ — различные символы из $\Sigma\setminus\{r\}$. Упорядочим, как и в доказательстве предыдущей теоремы, слова множества $(|\Sigma|\setminus\{r\})^{b_1}$, считая, что $a_i<a_j$, если $i<j$.

Докажем, что среди первых n слов из $(|\Sigma|\setminus\{r\})^{b_1}$ найдутся слова $\underbrace{a_1\ldots a_1}_{b_1}$ и $\underbrace{a_1\ldots a_1}_{b_1-1}a_2$. Слово $\underbrace{a_1\ldots a_1}_{b_1}$ является минимальным словом из $(|\Sigma|\setminus\{r\})^{b_1}$ при выбранном упорядочении, поэтому обязательно содержится среди первых n слов. Пусть $\underbrace{a_1\ldots a_1}_{b_1-1}a_2$ не содержится среди первых n слов. Тогда, так как слова меньше, чем $\underbrace{a_1\ldots a_1}_{b_1-1}a_2$, имеют вид $a_{i_1}\ldots a_{i_{b_1-1}}a_1$, то их всего $(|\Sigma|-1)^{b_1-1}$. По предположению первые n слов меньше слова $\underbrace{a_1\ldots a_1}_{b_1-1}a_2$, следовательно $n\leqslant(|\Sigma|-1)^{b_1-1}$. Но тогда $b_1=\left\lceil\log_{|\Sigma|-1}n\right\rceil\leqslant\left\lceil\log_{|\Sigma|-1}(|\Sigma|-1)^{b_1-1}\right\rceil=b_1-1$. Противоречие. Следовательно среди первых n слов имеются слова $\underbrace{a_1\ldots a_1}_{b_1}$ и $\underbrace{a_1\ldots a_1}_{b_1-1}a_2$. Обозначим первые n слов из $(|\Sigma|\setminus\{r\})^{b_1}$ как $\alpha_1,\ldots,\alpha_n$, причем без ограничения общности будем считать, что $\alpha_1=\underbrace{a_1\ldots a_1}_{b_1}$ и $\alpha_2=\underbrace{a_1\ldots a_1}_{b_1-1}a_2$. Для каждого $1\leqslant i\leqslant n$ в качестве выражений $\mathtt{R_{2i-1}}$ и $\mathtt{R_{2i}}$ возьмем выражения вида $\underbrace{\alpha_i.*\ \ldots\ .*\alpha_i}_{\frac{n_{2i-1}}{b_1}}$ и $\underbrace{\alpha_i.*\ \ldots\ .*\alpha_i}_{\frac{n_{2i}}{b_1}}$ соответственно.

Пусть $V_e = \langle Q_e, \Sigma, q_0^e, \delta_e, A_e \rangle$ — автомат приведенного вида, распознающий язык $L\left(.*(\mathtt{R_1}|\mathtt{R_3}).*(\mathtt{R_2}|\mathtt{R_4}).*\right)$. По построению из леммы 2 автомат V_e является тривиальным «слиянием» автоматов приведенного вида V_e' и V_e'', принимающих языки $L\left(.*(\mathtt{R_1}|\mathtt{R_3}).*\right)$ и $L\left(.*(\mathtt{R_2}|\mathtt{R_4}).*\right)$ соответственно. Заметим, что автоматы, распознающие языки $L\left(.*\mathtt{R_1}.*\right)$ и $L\left(.*\mathtt{R_3}.*\right)$, имеют такую же структуру, что и автоматы V_i из доказательства теоремы 2. Поэтому в автомате V_e' имеется $\dfrac{n_1}{b_1} \cdot \dfrac{n_3}{b_1}$ узловых состояний, которые достижимы и отличимы между собой. Причем если произвольное слово $\alpha \in \Sigma^*, \alpha \neq \alpha_1, \alpha \neq \alpha_2$ имеет длину не более чем b_1, то слово αr переводит любое узловое состояние в себя, а слова $\alpha_1 r$ и $\alpha_2 r$ переводят либо в другое узловое состояние, либо в финальное состояние. Аналогично автомат V_e'' имеет $\dfrac{n_2}{b_1} \cdot \dfrac{n_4}{b_1}$ узловых состояний. Заметим, что из доказательства достижимости узловых состояний теоремы 2 следует, что данные узловые состояния достигаются из начальных состояний словами, являющимися конкатенацией некоторого количества слов вида $\alpha_1 r$ и $\alpha_2 r$. Кроме того, финальные состояния автоматов V_e' и V_e'' не могут быть достижимы словами длины не более b_1, за исключением слов α_1 и α_2. Далее состояния автомата V_e, соответствующие узловым состояниям автоматов V_e' и V_e'', будем называть «почти узловыми» состояниями.

Пусть $V_i = \left\langle Q_i, \Sigma, q_0^i, \delta_i, A_i = \left\{q_f^i\right\} \right\rangle$ — автомат, принимающий язык $L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)$, где $i = \overline{1,n}$. Пронумеруем состояния автоматов в соответствии с их глубиной, считая, что начальные состояния имеют номер 0. Построим автомат V с $|V_e| \prod\limits_{i=3}^{n} |V_i|$ состояниями, где каждому состоянию ставится в соответствие набор состояний $\left(q_{j_e}^e, q_{j_3}^3, \ldots, q_{j_n}^n\right)$, где $q_{j_e}^e$ — состояние автомата V_e, а $q_{j_i}^i$ — состояния автоматов V_i соответственно. Для произвольного символа $a \in \Sigma$ зададим значение функции переходов как $\delta\left(\left(q_{j_e}^e, q_{j_3}^3, \ldots, q_{j_n}^n\right), a\right) = \left(\delta_e\left(q_{j_e}^e, a\right), \delta_3\left(q_{j_3}^3, a\right), \ldots, \delta_n\left(q_{j_n}^n, a\right)\right)$. В качестве начального состояния возьмем состояние $\left(q_0^e, q_0^3, \ldots, q_0^n\right)$, а в качестве финальных — все состояния вида $\left(q_{j_e}^e, q_{j_3}^3, \ldots, q_{j_n}^n\right)$, где либо найдется такое i, что $q_{j_i}^i = q_f^i$, либо состояние $q_{j_e}^e$ автомата V_e является финальным. Заметим, что по построению все финальные состояния являются поглощающими, а значит и неотличимыми. Такой автомат очевидно будет

принимать язык $L\,(.*(\mathtt{R_1}|\mathtt{R_3}).*(\mathtt{R_2}|\mathtt{R_4}).*) \cup \bigcup\limits_{i=3}^{n} L\,(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*)$, как автомат, реализующий по сути параллельную работу автомата V_e и $n-2$ автоматов V_i.

В данном доказательстве расширим определение узлового состояния. Назовем состояние вида $\left(q^e_{j_e}, q^3_{j_3}, \ldots, q^n_{j_n}\right)$ узловым, если $q^e_{j_e}$ соответствует почти узловому состоянию автомата V_e и $b_1 \mid j_i, q^i_{j_i} \neq q^i_f = q^i_{n_{2i-1}+n_{2i}}$ при $n = \overline{3,n}$. Докажем достижимость узловых состояний. Пусть достижимо состояние $\left(q^e_0, q^3_{j_3}, \ldots, q^n_{j_n}\right)$. Тогда, в силу того, что любое почти узловое состояние автомата V достигается из начального состояния q^e_0 конкатенацией некоторого числа слов вида $\alpha_1 r$ и $\alpha_2 r$, для любого почти узлового состояния q^e автомата V_e найдется слово $\alpha_{i_1} r \ldots \alpha_{i_k} r$, где $i_j = \overline{1,2}$, переводящее состояние q^e_0 в q^e. Заметим, что слова $\alpha_1 r$ и $\alpha_2 r$ переводят нефинальные состояния $q^i_{j_i}$, где $b_1 \mid j_i$, автоматов V_i в себя, а значит слово $\alpha_{i_1} r \ldots \alpha_{i_k} r$ переводит узловое состояние $\left(q^e_0, q^3_{j_3}, \ldots, q^n_{j_n}\right)$ в состояние $\left(q^e, q^3_{j_3}, \ldots, q^n_{j_n}\right)$. Докажем теперь индукцией по сумме индексов $\sum\limits_{i=3}^{n} j_i$, что достижимы узловые состояния вида $\left(q^e_0, q^3_{j_3}, \ldots, q^n_{j_n}\right)$. Состояние $\left(q^e_0, q^1_0, \ldots, q^n_0\right)$ является начальным, а следовательно достижимо. Пусть достижимы узловые состояния, у которых $\sum\limits_{i=3}^{n} j_i \leqslant k \cdot b_1$. Возьмем такое произвольное узловое состояние $\left(q^e_0, q^3_{j_3}, \ldots, q^n_{j_n}\right)$, что $\sum\limits_{i=3}^{n} j_i = (k+1) \cdot b_1$. Из определения узлового состояния следует, что найдется такое i, что $j_i \geqslant b_1$. По предположению состояние $\left(q^e_0, q^3_{j_3}, \ldots, q^i_{j_i - b_1}, \ldots, q^n_{j_n}\right)$ достижимо, так как сумма индексов состояний составляет $k \cdot b_1$. Слово $\alpha_i r$, очевидно, переводит состояние $q^i_{j_i - b_1}$ автомата V_i в состояние $q^i_{j_i}$, а остальные состояния $q^e_0, q^1_{j_1}, \ldots, q^n_{j_n}$ соответствующих автоматов переводит в себя. Следовательно, состояние $\left(q^1_{j_1}, \ldots, q^n_{j_n}\right)$ достижимо. Пусть достижимо состояние $\left(q^e_0, q^3_{j_3}, \ldots, q^n_{j_n}\right)$. Тогда, в силу того, что любое почти узловое состояние автомата V достигается из начального состояния q^e_0 конкатенацией некоторого числа слов вида $\alpha_1 r$ и $\alpha_2 r$, для любого почти узлового состояния q^e автомата V_e найдется слово $\alpha_{i_1} r \ldots \alpha_{i_k} r$, где $i_j = \overline{1,2}$, переводящее состояние q^e_0 в q^e. Заме-

тим, что слова $\alpha_1 r$ и $\alpha_2 r$ переводят нефинальные состояния $q^i_{j_i}$, где $b_1 \mid j_i$, автоматов V_i в себя, а значит слово $\alpha_{i_1} r \dots \alpha_{i_k} r$ переводит узловое состояние $\left(q^e_0, q^3_{j_3}, \dots, q^n_{j_n}\right)$ в состояние $\left(q^e, q^3_{j_3}, \dots, q^n_{j_n}\right)$. Докажем теперь индукцией по сумме индексов $\sum\limits_{i=3}^{n} j_i$, что достижимы узловые состояния вида $\left(q^e_0, q^3_{j_3}, \dots, q^n_{j_n}\right)$. Из этого следует, что все узловые состояния достижимы.

Докажем отличимость узловых состояний автомата V.

Пусть $\left(q^e_{j'_e}, q^3_{j'_3}, \dots, q^n_{j'_n}\right)$ и $\left(q^e_{j''_e}, q^3_{j''_3}, \dots, q^n_{j''_n}\right)$ — два различных узловых состояния. Пусть $q^e_{j'_e} \neq q^e_{q''_e}$. Так как данные состояния отличимы в автомате V_e, то найдется слово $\alpha_{i_1} r \dots \alpha_{i_k} r$, где $i_j = \overline{1,2}$, переводящее одно из них в финальное состояние, а другое — нет. Без ограничения общности будем считать, что в финальное состояние переводится $q^e_{j'_e}$. Тогда слово $\alpha_{i_1} r \dots \alpha_{i_k} r$ переводит состояние $\left(q^e_{j'_e}, q^3_{j'_3}, \dots, q^n_{j'_n}\right)$ в финальное состояние $\left(q^e_f, q^3_{j'_3}, \dots, q^n_{j'_n}\right)$, а состояние $\left(q^e_{j''_e}, q^3_{j''_3}, \dots, q^n_{j''_n}\right)$ в нефинальное состояние $\left(q^e, q^3_{j''_3}, \dots, q^n_{j''_n}\right)$, где q^e — некоторое нефинальное состояние автомата V_e. Пусть $q^e_{j'_e} = q^e_{q''_e}$. Без ограничения общности будем считать, что $j'_3 > j''_3$. Тогда, очевидно, слово $\underbrace{\alpha_1 r \dots \alpha_1 r}_{\frac{n_5+n_6-j'_3}{b_1}}$ переводит состояние $\left(q^e_{j'_e}, q^3_{j'_3}, \dots, q^n_{j'_n}\right)$ в состояние $\left(q^e_{j'_e}, q^3_{n_5+n_6}, q^4_{j'_4} \dots, q^n_{j'_n}\right)$, а состояние $\left(q^e_{j''_e}, q^3_{j''_3}, \dots, q^n_{j''_n}\right)$ — в состояние $\left(q^e_{j''_e}, q^3_{j''_3+n_5+n_6-j'_3}, q^4_{j''_4} \dots, q^n_{j''_n}\right)$. По построению первое состояние является финальным, а второе — нет, так как $n_5 + n_6 + j''_3 - j'_3 < n_5 + n_6$. Таким образом автомат содержит не менее $\left(\frac{n_1}{b_1} \cdot \frac{n_3}{b_1} + \frac{n_2}{b_1} \cdot \frac{n_4}{b_1}\right) \cdot \prod\limits_{i=3}^{n} \left(\frac{n_{2i-1}}{b_1} + \frac{n_{2i}}{b_1}\right)$ узловых состояний.

Пусть $b_1 = 1$. Тогда, так как помимо узловых состояний, автомат содержит одно финальное состояние, утверждение (9) верно.

Пусть $b_1 > 1$. Рассмотрим отдельно узловые состояния, которые не могут быть переведены в финальное состояние словом длины b_1. Докажем, что словами длины не более $b_1 - 1$ из каждого такого состояния достижимо не менее $\sum_{i=1}^{b_1-1} (|\Sigma| - 1)^i$ состояний отличимых друг

от друга и от узловых состояний. Пусть $q = \left(q^e_{j_e}, q^3_{j_3}, \ldots, q^n_{j_n}\right)$ — некоторое такое узловое состояние. Далее через $q_\alpha = \left(q^e_\alpha, q^3_\alpha, \ldots, q^n_\alpha\right)$ будем обозначать состояние, в которое по слову α переходит состояние q. Отметим, что другие узловые состояния не достижимы словами длины меньше b_1. Докажем, что состояния вида q_α, где α — слово длины не более b_1-1, отличимы. В соответствии с выбором b_1 и слов α_i, для любого слова $\alpha \in (\Sigma \setminus \{r\})^{b_1-1}$ найдется такое $1 \leqslant i \leqslant n$, что $\alpha a_1 = \alpha_i$. Возьмем $q_{\alpha'}$ и $q_{\alpha''}$ — два из рассматриваемых состояния, причем $\alpha' \neq \alpha''$. Пусть длины α' и α'' совпадают. Без ограничения общности будем считать, что $\alpha' \neq a_1 \ldots a_1$. Как было отмечено выше, найдется такое i', что $\alpha_{i'} = \alpha'\beta'$, где β' — непустое слово из $(\Sigma \setminus \{r\})^*$ длины $b_1 - |\alpha'|$. Причем, так как $\alpha' \neq a_1 \ldots a_1$, $i' \geqslant 3$. Тогда, так как слово $\alpha_{i'}r$, очевидно, переводит состояние $q = \left(q^e_{j_e}, q^3_{j_3}, \ldots, q^{i'}_{j_i+b_1}, \ldots, q^n_{j_n}\right)$ в узловое состояние $\left(q^e_{j_e}, q^3_{j_3}, \ldots, q^{i'}_{j_{i'}+b_1}, \ldots, q^n_{j_n}\right)$, слово $\beta'r$ также переводит $q_{\alpha'}$ в это узловое состояние. Пусть найдется такое i'', что $\alpha_{i''} = \alpha''\beta'$. Если $i'' \geqslant 3$, то в силу того, что $\alpha' \neq \alpha''$, слово $\alpha_{i'}$ также не совпадает с $\alpha_{i''}$. Аналогично предыдущим рассуждениям слово $\alpha_{i''}r$ переводит q в $\left(q^e_{j_e}, q^3_{j_3}, \ldots, q^{i''}_{j_{i''}+b_1}, \ldots, q^n_{j_n}\right)$, а $\beta'r$ переводит состояние $q_{\alpha''}$ в него же. Если $i'' \leqslant 2$, то слово $\alpha_{i''}r$ переводит состояние q в отличное от него узловое состояние $\left(q^e_{j''_e}, q^3_{j_3}, \ldots, q^n_{j_n}\right)$, а $\beta'r$ переводит состояние $q_{\alpha''}$ в него же. Таким образом в обоих случаях $\beta'r$ переводит состояния $q_{\alpha'}$ и $q_{\alpha''}$ в отличимые узловые состояния. Следовательно они также отличимы. Если не найдется такое i'', что $\alpha_{i''} = \alpha''\beta'$, то $\beta'r$ переводит $q_{\alpha''}$ в состояние q. Узловые состояния $\left(q^1_{j_1}, \ldots, q^{i'}_{j_{i'}+b_1}, \ldots, q^n_{j_n}\right)$ и q отличимы, а значит состояния $q_{\alpha'}$ и $q_{\alpha''}$, переводимые в них словом $\beta'r$, также отличимы.

Пусть теперь длины α' и α'' отличаются. Без ограничения общности будем считать, что длина α' больше. Найдется такое i, что $\alpha_i = \alpha'\beta'$, где β' — непустое слово из $(\Sigma \setminus \{r\})^*$ длины $b_1 - |\alpha'|$. Аналогично предыдущим рассуждениям $\beta'r$ переводит $q_{\alpha'}$ либо в узловое состояние $\left(q^e_{j_e}, q^3_{j_3}, \ldots, q^i_{j_i+b_1}, \ldots, q^n_{j_n}\right)$, если $i \geqslant 3$, либо, в противном случае, в узловое состояние $\left(q^e_{j'_e}, q^3_{j_3}, \ldots, q^n_{j_n}\right)$, где j'_e — почти узловое состояние автомата V_e, в которое переводится словом α_i состояние $q^e_{j_e}$. Причем оба этих узловых состояния отличны от состояния

q. При этом слово $\alpha''\beta' r$ переводит состояние q в себя, так как по предположению длина $\alpha''\beta'$ меньше b_1. Узловые состояния, в которые переводятся словом $\beta' r$ состояния $q_{\alpha'}$ и $q_{\alpha''}$, отличимы, а значит $q_{\alpha'}$ и $q_{\alpha''}$ также отличимы.

Заметим, что рассматриваемые состояния, достижимые словами длины не более b_1-1 из различных узловых состояний, отличимы, так как переход по символу r переводит их в соответствующие узловые состояния, которые по доказанному ранее отличимы.

По построению автомата V_e'' и из доказательства теоремы 2 следует, что $\left(\frac{n_2}{b_1}-1\right)\cdot\left(\frac{n_4}{b_1}-1\right)$ узловых состояния автомата V_e'' не могут перейти в финальное состояние словом длины b_1. Следовательно автомат V_e имеет $\left(\frac{n_1}{b_1}\cdot\frac{n_3}{b_1}+\left(\frac{n_2}{b_1}-1\right)\cdot\left(\frac{n_4}{b_1}-1\right)\right)$ почти узловых состояний, которые не могут перейти в финальное состояние. Таким образом число состояний автомата приведенного вида, распознающего язык $L\left(.*\left(\mathtt{R_1}|\mathtt{R_3}\right).*\left(\mathtt{R_2}|\mathtt{R_4}\right).*\right)\cup\bigcup\limits_{i=3}^{n}L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)$ с выражениями R_j, заданными выше, не меньше, чем

$$\left(\frac{n_1}{b_1}\cdot\frac{n_3}{b_1}+\frac{n_2}{b_1}\cdot\frac{n_4}{b_1}\right)\cdot\prod_{i=3}^{n}\left(\frac{n_{2i-1}}{b_1}+\frac{n_{2i}}{b_1}\right)+\left(\sum_{i=1}^{b_1-1}\left(|\Sigma|-1\right)^i\right)\times$$
$$\times\left(\frac{n_1}{b_1}\cdot\frac{n_3}{b_1}+\left(\frac{n_2}{b_1}-1\right)\cdot\left(\frac{n_4}{b_1}-1\right)\right)\cdot\prod_{i=3}^{n}\left(\frac{n_{2i-1}}{b_1}+\frac{n_{2i}}{b_1}-1\right)+1,$$

где первое слагаемое — число узловых состояний, второе — число неузловых состояний, соответствующих $\left(\frac{n_1}{b_1}\cdot\frac{n_3}{b_1}+\left(\frac{n_2}{b_1}-1\right)\cdot\left(\frac{n_4}{b_1}-1\right)\right)\cdot\prod\limits_{i=3}^{n}\left(\frac{n_{2i-1}}{b_1}+\frac{n_{2i}}{b_1}-1\right)$ узловым состояниям, из которых нельзя попасть в финальное словом длины b_1, а последнее слагаемое — одно финальное состояние.

Докажем оценку (10). Пусть a — некоторый символ алфавита Σ. Введем следующие обозначения: $\Sigma'=\Sigma\setminus\{a\}$, $b_1'=\left\lceil\log_{|\Sigma'|-1}n\right\rceil=\left\lceil\log_{|\Sigma|-2}n\right\rceil=b_2$ и $n_i'=\left\lfloor\frac{n_i}{b_1'}\right\rfloor\cdot b_1'$. Тогда, согласно доказательству утверждения (9), найдутся такие выражения $\mathtt{R'_j}$, $j=\overline{1,2n}$ над алфавитом Σ', что $\left|V\left(L^{pf}\left(.*\mathtt{R'_i}\right)\right)\right|=n_i'+2$ и

$$\left|V\left(L\left(.*(\mathtt{R_1}|\mathtt{R_3}).*(\mathtt{R_2}|\mathtt{R_4}).*\right)\cup\bigcup_{i=3}^{n}L\left(.*\mathtt{R'_{2i-1}}.*\mathtt{R'_{2i}}.*\right)\right)\right|\geqslant$$

$$\geqslant\left(\frac{n'_1}{b'_1}\cdot\frac{n'_3}{b'_1}+\frac{n'_2}{b'_1}\cdot\frac{n'_4}{b'_1}\right)\cdot\prod_{i=3}^{n}\left(\frac{n'_{2i-1}}{b'_1}+\frac{n'_{2i}}{b'_1}\right)+\left(\sum_{i=1}^{b'_1-1}\left(\left|\Sigma'\right|-1\right)^i\right)\times$$

$$\times\left(\frac{n'_1}{b'_1}\cdot\frac{n'_3}{b'_1}+\left(\frac{n'_2}{b'_1}-1\right)\cdot\left(\frac{n'_4}{b'_1}-1\right)\right)\cdot\prod_{i=3}^{n}\left(\frac{n'_{2i-1}}{b'_1}+\frac{n'_{2i}}{b'_1}-1\right)+1.$$

В качестве выражений $\mathtt{R_i}$ возьмем выражения $\mathtt{R'_i}$, если $n_i = n'_i$, и $(\mathtt{R'_i}|a\{n_i - n'_i + 1\})$ в противном случае. Отметим, что выражение вида $\mathtt{R'}.*\mathtt{R''}$ над алфавитом $\Sigma\setminus\{a\}$ совпадает с выражением $\mathtt{R'}[\mathtt{c_1}\ldots\mathtt{c_{|\Sigma|-1}}]\mathtt{R''}$, где $c_1,\ldots,c_{|\Sigma|-1}$ — различные символы алфавита Σ, отличные от символа a. Таким образом, выражения вида $\mathtt{R'}.*\mathtt{R''}.*\ldots.*\mathtt{R}^{(\mathtt{k})}$ над алфавитом $\Sigma\setminus\{a\}$ при переходе к алфавиту Σ примут вид $\mathtt{R'}[\hat{}\,a]*\mathtt{R''}[\hat{}\,a]*\ldots[\hat{}\,a]*\mathtt{R}^{(\mathtt{k})}$. По лемме 6 число состояний автомата $V\left(L\left(.*\mathtt{R'_i}.*\right)\right)$ над алфавитом Σ совпадает с числом состояний автомата $V\left(L\left(.*\mathtt{R'_i}.*\right)\right)$ над алфавитом Σ'. Применив следствие 1, получаем, что $\left|V\left(L^{pf}\left(.*\mathtt{R'_i}\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R'_i}.*\right)\right)\right|+1 = n'_i+2$, где $V\left(L^{pf}\left(.*\mathtt{R'_i}\right)\right)$ — автомат над алфавитом Σ. При $n_i = n'_i$ очевидно, что $\left|V\left(L^{pf}\left(.*\mathtt{R_i}\right)\right)\right| = \left|V\left(L^{pf}\left(.*\mathtt{R'_i}\right)\right)\right| = n'_i + 2$. В случае $n_i \neq n'_i$ по лемме 4, в силу того, что выражение $\mathtt{R'_i}$ определено над алфавитом $\Sigma\setminus\{a\}$, верно равенство $\left|V\left(L\left(.*(\mathtt{R'_i}|a\{n_i - n'_i + 1\}).*\right)\right)\right| = \left|V\left(L\left(.*\mathtt{R'_i}.*\right)\right)\right| + \left|V\left(L\left(.*a\{n_i - n'_i + 1\}.*\right)\right)\right| - 2$. Очевидно, что $\left|V\left(L\left(.*a\{n_i - n'_i + 1\}.*\right)\right)\right| = n_i - n'_i + 2$. Поэтому, применив следствие 1, получаем: $\left|V\left(L^{pf}\left(.*\mathtt{R_i}\right)\right)\right| = \left|V\left(L\left(.*(\mathtt{R'_i}|a\{n_i - n'_i + 1\}).*\right)\right)\right| + 1 = \left|V\left(L^{pf}\left(.*\mathtt{R'_i}\right)\right)\right| - 1 + n_i - n'_i + 2 - 2 + 1 = n'_i + n_i - n'_i + 2 = n_i + 2$.

Заметим, что, как и в теореме 2, регулярный язык, принимаемый автоматом $V' = V\left(L\left(.*(\mathtt{R'_1}|\mathtt{R'_3}).*(\mathtt{R'_2}|\mathtt{R'_4}).*\right)\cup\bigcup\limits_{i=3}^{n}L\left(.*\mathtt{R'_{2i-1}}.*\mathtt{R'_{2i}}.*\right)\right)$ над алфавитом Σ', совпадает с языком $\left(\bigcup\limits_{i=1}^{n}L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)\right)\cap(\Sigma')^*$. Поэтому по лемме 5

$$\left|V\left(\bigcup_{i=1}^{n}L\left(.*\mathtt{R_{2i-1}}.*\mathtt{R_{2i}}.*\right)\right)\right|\geqslant\left|V'\right|\geqslant\left(\frac{n'_1}{b'_1}\cdot\frac{n'_3}{b'_1}+\frac{n'_2}{b'_1}\cdot\frac{n'_4}{b'_1}\right)\times$$

$$\times \prod_{i=3}^{n}\left(\frac{n'_{2i-1}}{b'_1}+\frac{n'_{2i}}{b'_1}\right)+\left(\sum_{i=1}^{b'_1-1}\left(|\Sigma'|-1\right)^i\right)\times$$

$$\times\left(\frac{n'_1}{b'_1}\cdot\frac{n'_3}{b'_1}+\left(\frac{n'_2}{b'_1}-1\right)\cdot\left(\frac{n'_4}{b'_1}-1\right)\right)\cdot\prod_{i=3}^{n}\left(\frac{n'_{2i-1}}{b'_1}+\frac{n'_{2i}}{b'_1}-1\right)+1=$$

$$=\left(\left\lfloor\frac{n_1}{b_2}\right\rfloor\cdot\left\lfloor\frac{n_3}{b_2}\right\rfloor+\left\lfloor\frac{n_2}{b_2}\right\rfloor\cdot\left\lfloor\frac{n_4}{b_2}\right\rfloor\right)\cdot\prod_{i=3}^{n}\left(\left\lfloor\frac{n_{2i-1}}{b_2}\right\rfloor+\left\lfloor\frac{n_{2i}}{b_2}\right\rfloor\right)+$$

$$+\left(\sum_{i=1}^{b_2-1}\left(|\Sigma|-2\right)^i\right)\cdot\left(\left\lfloor\frac{n_1}{b_2}\right\rfloor\cdot\left\lfloor\frac{n_3}{b_2}\right\rfloor+\left(\left\lfloor\frac{n_2}{b_2}\right\rfloor-1\right)\cdot\left(\left\lfloor\frac{n_4}{b_2}\right\rfloor-1\right)\right)\cdot$$

$$\cdot\prod_{i=3}^{n}\left(\left\lfloor\frac{n_{2i-1}}{b_2}\right\rfloor+\left\lfloor\frac{n_{2i}}{b_2}\right\rfloor-1\right)+1.$$

Доказательство теоремы 4

Согласно утверждению (4) теоремы 1 для любого $C\in(0;1]\cap\mathbb{Q}$ найдутся такие выражения $\mathtt{R}_1$, $\mathtt{R}_2$, $\mathtt{R}_3$ и $\mathtt{R}_4$, что

$$C=\frac{|V\left(L\left(.*\left(\mathtt{R}'_1|\mathtt{R}'_3\right).*\left(\mathtt{R}'_2|\mathtt{R}'_4\right).*\right)\right)|}{|V\left(L\left(.*\mathtt{R}'_1.*\mathtt{R}'_2.*\right)\cup L\left(.*\mathtt{R}'_3.*\mathtt{R}'_4.*\right)\right)|}.$$

Рассмотрим выражения $\mathtt{R}^1=.*\mathtt{R}'_1.*\mathtt{R}'_2.*$, $\mathtt{R}^j=.*\mathtt{R}'_3.*\mathtt{R}_4.*$, $j=\overline{2,n}$. Заметим, что $L\left(.*\left(\mathtt{R}_1|\mathtt{R}_3\right).*\left(\mathtt{R}_2|\mathtt{R}_4\right).*\right)\supseteq L\left(.*\mathtt{R}_3.*\mathtt{R}_4.*\right)$. Тогда

$$\frac{\left|V\left(L\left(.*\left(\mathtt{R}_1|\mathtt{R}_3\right).*\left(\mathtt{R}_2|\mathtt{R}_4\right).*\right)\cup\bigcup\limits_{i=3}^{n}L\left(\mathtt{R}^i\right)\right)\right|}{\left|V\left(\bigcup\limits_{i=1}^{n}L\left(\mathtt{R}^i\right)\right)\right|}=$$

$$=\frac{|V\left(L\left(.*\left(\mathtt{R}_1|\mathtt{R}_3\right).*\left(\mathtt{R}_2|\mathtt{R}_4\right).*\right)\right)|}{|V\left(L\left(\mathtt{R}^1\right)\cup L\left(\mathtt{R}^2\right)\right)|}=C.$$

Применение алгоритма

Для работы с регулярными выражениями и конечными автоматами был разработан программный комплекс [15], в котором реализована возможность модификации выражений вида $.*\mathtt{R}_1.*\mathtt{R}_2.*$.

Для проверки эффективности модификации регулярных выражений была выбрана база сигнатур сетевой системы обнаружения вторжений Snort [16]. Из нее были взяты 11 выражений вида $\mathtt{R}^{\mathtt{i}} = .*\mathtt{R}'_{\mathtt{i}}.*\mathtt{R}''_{\mathtt{i}}.*$. Затем был построен автомат $V\left(\bigcup\limits_{i=1}^{11} L\left(\mathtt{R}^{\mathtt{i}}\right)\right)$ и для каждой пары выражений $\mathtt{R}^{\mathtt{i}}$ и $\mathtt{R}^{\mathtt{j}}$ $(i \neq j)$ был построен автомат $V\left(L\left(.*(\mathtt{R}'_{\mathtt{i}}|\mathtt{R}'_{\mathtt{j}}).*(\mathtt{R}''_{\mathtt{i}}|\mathtt{R}''_{\mathtt{j}}).*\right) \bigcup\limits_{k\neq i,j} L\left(\mathtt{R}^{\mathtt{k}}\right)\right)$. Гистограмма значений отношения $C = \dfrac{\left|V\left(L\left(.*(\mathtt{R}'_{\mathtt{i}}|\mathtt{R}'_{\mathtt{j}}).*(\mathtt{R}''_{\mathtt{i}}|\mathtt{R}''_{\mathtt{j}}).*\right) \bigcup\limits_{k\neq i,j} L\left(\mathtt{R}^{\mathtt{k}}\right)\right)\right|}{\left|V\left(\bigcup\limits_{i=1}^{11} L\left(\mathtt{R}^{\mathtt{i}}\right)\right)\right|}$, характеризующего эффективность алгоритма в случае набора из более чем двух выражений, изображена на рисунке 5.

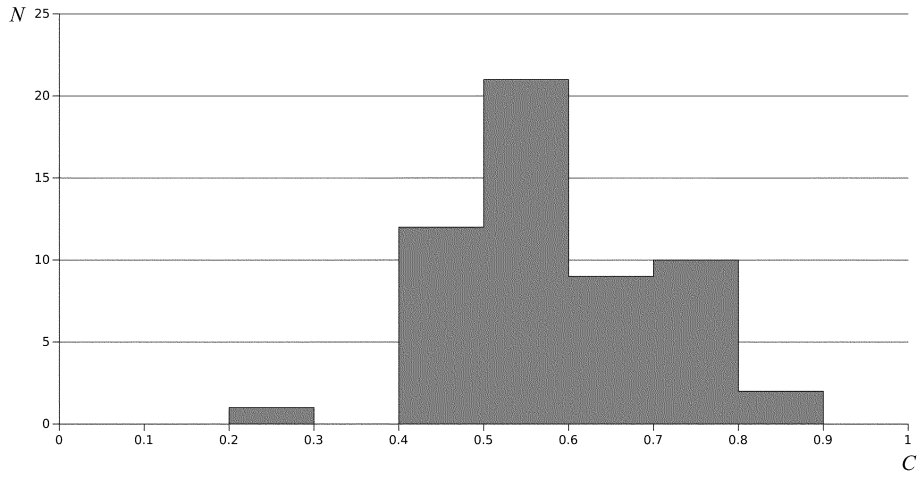

Рис. 5. Гистограмма эффективности алгоритма.

Видно, что применение данного алгоритма изменения выражений для различных пар выражений в большинстве случаев привело к сокращению числа состояний до 50–70%. При этом в одном случаев число состояний сократилось до примерно 25%. Таким образом модификация может быть эффективна, но это существенно зависит от выбора пары выражений для изменения.

Заключение

В статье было рассмотрено расширение метода изменения регулярных выражений [11] для сокращения числа состояний детерминированных конечных автоматов, реализующих поиск по выражениям. Были даны оценки на число состояний автомата при таком изменении в случае алфавита, состоящего из не менее чем трех символов. Также доказано, что относительное уменьшение числа состояний может быть произвольным.

Автор выражает благодарность к.ф.-м.н. Галатенко Алексею Владимировичу и к.ф.-м.н. Панкратьеву Антону Евгеньевичу за постановку задачи и внимание к работе, Александрову Денису Евгеньевичу за помощь в редактировании статьи.

Список литературы

[1] Документация системы Snort. http://www.snort.org/documents.

[2] Документация системы Bro. http://www.bro.org/.

[3] Описание системы L7-filter. http://l7-filter.sourceforge.net/README.

[4] Аппаратные продукты компании Cisco. http://www.cisco.com/c/en/us/products/security/intrusion-prevention-system-ips.

[5] Kumar S. et al. Algorithms to accelerate multiple regular expressions matching for deep packet inspection // ACM SIGCOMM Computer Communication Review. — ACM, 2006. — Т. 36. № 4. — P. 339–350.

[6] Aho A. V., Corasick M. J. Efficient string matching: an aid to bibliographic search // Communications of the ACM. — 1975. — Т. 18. № 6. — P. 333–340.

[7] Liu C., Wu J. Fast Deep Packet Inspection with a Dual Finite Automata // Computers, IEEE Transactions on. — 2013. — Т. 62. № 2. — P. 310–321.

[8] Kumar S. et al. Curing regular expressions matching algorithms from insomnia, amnesia, and acalculia // Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems. — ACM, 2007. — P. 155–164.

[9] Smith R. et al. Deflating the big bang: fast and scalable deep packet inspection with extended finite automata // ACM SIGCOMM Computer Communication Review. — ACM, 2008. — Т. 38. № 4. — P. 207–218.

[10] Yu F. et al. Fast and memory-efficient regular expression matching for deep packet inspection // Architecture for Networking and Communications systems, ANCS 2006. ACM/IEEE Symposium on. — IEEE, 2006. — P. 93–102.

[11] Александров Д. Е. Об уменьшении автоматной сложности за счет расширения регулярных языков // Программная инженерия. — 2014. — № 11. — С. 26–34

[12] Регулярные выражения PCRE. http://www.pcre.org/pcre.txt.

[13] Кудрявцев В. Б., Алешин С. В., Подколзин А. С. Введение в теорию автоматов. — М.: Наука, 1985.

[14] Martin J. C. Introduction to Languages and the Theory of Computation. — Изд. 4-е. — New York: McGraw-Hill, 2011.

[15] Программный комплекс RE2FA / Александров Д. Е. Свидетельство о государственной регистрации программы для ЭВМ № 2014614857. — 2014.

[16] База сигнатур системы Snort. http://www.snort.org/snort-rules/.