Секция «Дискретная математика и математическая кибернетика»

Об оценках мощности некоторых классов регулярных языков

Александров Дмитрий Евгеньевич

Аспирант

Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра математической теории
интеллектуальных систем, Москва, Россия

E-mail: dalexandrov@intsys.msu.ru

С развитием технологий передачи данных одним из актуальных направлений информационной безопасности стали исследование и фильтрация сетевого трафика. На настоящее время существует большое число различных сетевых систем обнаружения вторжений, таких как Snort, Bro, L7-filter и аппаратные продукты фирмы Cisco, базы сигнатур которых представляют собой наборы регулярных выражений. Вердикт о вредоносности трафика в таких системах выносится на основании соответствия фильтруемых данных хотя бы одному из регулярных выражений базы.

Традиционно для проверки принадлежности слова регулярному языку, задаваемому набором регулярных выражений, используются *детерминированные конечные автоматы* (ДКА) [1]. Данный метод характеризуется крайне низкой сложностью вычисления. Однако с ростом числа распознаваемых выражений увеличивается пространственная сложность — число состояний распознающего ДКА, и, соответственно, требуемый для программной реализации алгоритма объем памяти. В случае, когда число состояний автомата экспоненциально зависит от количества регулярных выражений, говорят о так называемом “экспоненциальном взрыве”. Среди различных видов выражений, приводящих к экспоненциальному взрыву, можно выделить класс выражений вида $.*\mathtt{R}_1.*\mathtt{R}_2.*$, где $\mathtt{R}_1$ и $\mathtt{R}_2$ — регулярные выражения, а подвыражения $.*$ задают регулярные языки, совпадающие со множеством всех слов Σ^*, где Σ — алфавит, над которым заданы выражения. Отметим, что выражения такого вида часто встречаются на практике в системах обеспечения информационной безопасности. Например, база сигнатур сетевой системы обнаружения вторжений Snort [2] содержит 36 таких выражений, причем детерминированный конечный автомат, распознающий регулярный язык, задаваемый только 11 из 36 выражений, содержит более 1,5 млн. состояний.

Предлагаемый в работе [3] подход к преодолению проблемы экспоненциального взрыва для случая выражений вида $.*\mathtt{R}_1.*\mathtt{R}_2.*$ предполагает модификацию исходного набора регулярных выражений для сокращения числа состояний распознающего ДКА.

Пусть заданы два регулярных выражения $\mathtt{R}^1 = .*\mathtt{R}_1.*\mathtt{R}_2.*$ и $\mathtt{R}^2 = .*\mathtt{R}_3.*\mathtt{R}_4.*$. Для снижения пространственной сложности ДКА, распознающего принадлежность слова регулярному языку, который задается парой выражений $\mathtt{R}^1$ и $\mathtt{R}^2$, предлагается заменить исходные выражения $\mathtt{R}^1$ и $\mathtt{R}^2$ на одно выражение $.*(\mathtt{R}_1|\mathtt{R}_3).*(\mathtt{R}_2|\mathtt{R}_4).*$. С одной стороны, такое изменение гарантирует снижение числа состояний распознающего ДКА. С другой стороны, хотя модификация выражений приводит к изменению распознаваемого регулярного языка, новый язык полностью содержит исходный регулярный язык. Следовательно, при распознавании принадлежности слова исходному регулярному языку возможна только ошибка первого рода — ложное срабатывание, а, значит, данный метод применим в системах информационной безопасности, для которых недопустимы ошибки второго рода при распознавании.

В работах [3,4] доказаны оценки на число состояний автомата при такой модификации выражений в случае наборов из двух и более выражений.

Обозначим через $L(\mathtt{R})$ регулярный язык, определяемый регулярным выражением $\mathtt{R}$. Через $L^{pf}(\mathtt{R})$ обозначим такое наибольшее подмножество $L(\mathtt{R})$, что ни одно из слов из $L(\mathtt{R})$ не является нетривиальным префиксом для слов из $L^{pf}(\mathtt{R})$. Под нетривиальным префиксом в данном случае понимаем префикс, не совпадающий со всем словом. В случае если $L(\mathtt{R})$ содержит пустую строку Λ, положим $L^{pf}(\mathtt{R}) = \{\Lambda\}$. Причем нетрудно показать, что $L^{pf}(\mathtt{R})$ всегда является регулярным языком.

Пусть $G_l(L) = |\{\alpha \in L \mid |\alpha| = l\}|$ — функция роста непустого языка $L \subseteq \Sigma^*$, тогда через $C_l(L)$ обозначим отношение $\dfrac{G_l(L)}{|\Sigma|^{l-m} \cdot G_m(L)}$, где $m = \min\limits_{\alpha \in L} |\alpha|$.

Обозначим через RF множество таких регулярных выражений $\mathtt{R}$, что: $L(\mathtt{R})$ не содержит пустого слова; $\forall \alpha \in L(\mathtt{R})$ никакое слово из $L(\mathtt{R})$, кроме α, не содержит подслово α; $\forall \alpha, \beta \in L(\mathtt{R})$ никакой нетривиальный суффикс α не является префиксом β.

Пусть регулярное выражение $\mathtt{R}$ принадлежит классу выражений RF и имеет вид $\mathtt{R_1c_1{*}R_2c_2{*}}\ldots$, тогда через $C^{max}(L(.{*}\mathtt{R}))$ обозначим произведение $\left(\prod\limits_{i=1}^{p+1} \sum\limits_{j=m_i}^{\infty} C_j(\mathtt{R_i})\right) \cdot \left(\prod\limits_{i=1}^{p} \dfrac{|\Sigma|}{|\Sigma| - k_i}\right)$, где $m_i = \min\limits_{\alpha \in L(\mathtt{R_i})} |\alpha|$, $k_i = |L(\mathtt{c_i})|$.

В работе [5] предложен способ оценки относительного роста числа слов регулярного языка, задаваемого парой регулярных выражений, при модификации данных выражений.

Теорема. Пусть $.{*}\mathtt{R_1}.{*}\mathtt{R_2}.{*}$ и $.{*}\mathtt{R_3}.{*}\mathtt{R_4}.{*}$ — такие регулярные выражения, что подвыражения $\mathtt{R_i}$ принадлежат классу RF и имеют вид $\mathtt{R_1^ic_1^i{*}R_2^ic_2^i{*}}\ldots\mathtt{R_{p_i}^i}$ и $m_1 \geqslant m_3$, $m_2 \geqslant m_4$, где $m_i = \min\limits_{\alpha \in L(\mathtt{R_i})} |\alpha|$. Пусть $0 < \varepsilon < 1$, тогда для всех таких длин l, что $m_3 + m_4 \leqslant l \leqslant \dfrac{\varepsilon}{\frac{G_{m_i}(L(\mathtt{R_i}))}{|\Sigma|^{m_i}} \cdot C^{max}(L^p}$ при $i \in \{3, 4\}$, верно неравенство:

$$\frac{G_l(L(.{*}(\mathtt{R_1}|\mathtt{R_3}).{*}(\mathtt{R_2}|\mathtt{R_4}).{*}))}{G_l(L(.{*}\mathtt{R_1}.{*}\mathtt{R_2}.{*}) \cup L(.{*}\mathtt{R_3}.{*}\mathtt{R_4}.{*}))} \leqslant 1 + \frac{1}{1-\varepsilon} \cdot \left(\frac{C^{max}\left(L^{pf}(.{*}\mathtt{R_1})\right) \cdot G_{m_1}(L(\mathtt{R_1}))}{|\Sigma|^{m_1-m_3} \cdot G_{m_3}(L(\mathtt{R_3}))} + \frac{C^{max}\left(L^{pf}(.{*}\mathtt{R_2})\right) \cdot G_{m_2}(L(\mathtt{R_2}))}{|\Sigma|^{m_2-m_4} \cdot G_{m_4}(L(\mathtt{R_4}))} \right).$$

Такой подход позволяет оценить рост языка при модификации выражений сразу для некоторого промежутка значений длин слов.

На практике (алфавит Σ содержит 256 символов) значения произведений $\dfrac{G_{m_i}(L(\mathtt{R_i}))}{|\Sigma|^{m_i}} \cdot C^{max}\left(L^{pf}(.{*}\mathtt{R_i})\right)$ близки к $\dfrac{G_{m_i}(L(\mathtt{R_i}))}{|\Sigma|^{m_i}} \cdot \prod\limits_{j=1}^{p} \dfrac{|\Sigma|}{|\Sigma| - k_j}$ и, в случае выражений системы Snort, в большинстве случаев имеют порядок 10^{-8} и меньше, то есть, например, при $\varepsilon = 10^{-2}$ максимальная длина слов, на которых верна оценка, имеет порядок свыше 10^6. Верхняя же оценка относительного роста регулярного языка, предложенная в теореме выше, для пар выражений системы Snort в более чем трети случаев (при $\varepsilon = 10^{-2}$) не превышает 10^{-2}.

Источники и литература

1) Кудрявцев В. Б., Алешин С. В., Подколзин А. С. Введение в теорию автоматов. Москва.: Наука. Гл. ред. физ.-мат. лит., 1985.
2) Snort.Org [Электронный ресурс] База сигнатур системы Snort – Электрон. дан. – Режим доступа: https://www.snort.org/downloads/#rule-downloads, свободный. – Загл. с экрана.
3) Александров Д.Е. Об уменьшении автоматной сложности за счет расширения регулярных языков // Программная инженерия. 2014. № 11. С. 26–34.
4) Александров Д.Е. Об оценках автоматной сложности распознавания класса регулярных языков // Интеллектуальные системы. 2014. Т. 18. № 4. С. 121–146.
5) Александров Д.Е. Об оценках мощности некоторых классов регулярных языков // Дискретная математика. 2015 (в печати).

Слова благодарности

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.