Пересечение контекстно свободного и регулярного языков

Арсений Терехов

September 2019

1 Постановка задачам

Пусть есть cfg $G = (V_n, V_t, P, S)$ и pda $N = (Q, A, P, q_0, F)$. Тогда мы хотим найти кс грамматику $G : L(G) = L(G) \cap L(N)$.

2 Что мы знаем

Любая кс грамматика представима в нормальной форме Хомского или Грейбах. Так же для любой cfg есть npda (недетерминированный магазинный автомат), и наоборот.

3 Идея и мысли в слух

Если бы у нас был сразу npda, и нам надо было вернуть новый npda, тогда всё было бы гораздо проще, потому что пересечение двух pda легко делается декартовым произведением их множеств состояний (правила для стека тоже легко добавить).

Но на вход нам поступает именно cfg. С одной стороны, её легко превратить в npda, приведя к нормальной форме Грейбах, с другой стороны, преобразовывать npda к cfg не хочется. Так что давайте не будем ничего никуда переводить, а сразу строить новую cfg.

Пусть имеется правило $A \to a_1, a_2, ..., a_n$, где $a_i \in V_n \cup V_t$. Давайте каждый элемент из правой части окружим состояниями dfa и получим нетерминалы вида $[q_i, a_j, q_k]$ (то есть вся эта тройка - один нетерминал). Тогда каждое правило cfg (где A не является стартовым нетерминалом S) $A \to a_1, a_2, ..., a_n$ мы заменяем на множество правил $\{[q_1, A, q_{n+1}] \to [q_1, a_1, q_2], [q_2, a_2, q_3], ..., [q_n, a_n, q_{n+1}] \ : \ (q_1, q_2, ..., q_{n+1}) \in Qn + 1\}$. Для начальной вершины мы добавляем правила вида $S \to [q_0, S, q_f] : q_f \in F$.

Давайте теперь добавим правила для терминалов - для каждого правила из dfa $(q_1, a) \to q_2$ добавим правила вида $[q_1, a, q_2] \to a$. Таким образом новыми терминалами построенной грамматики будет алфавит dfa.

4 Про корректность

Вообще, хорошо бы формально показать равенство множеств. Но я сильно ограничен по времени и совсем не успеваю, простите) Вообще мы на самом деле внедрили dfa в нетерминалы грамматики. Так что если посмотреть на вывод слова в новой грамматике, самые глубокие нетерминалы в дереве вывода образуют вывод слова в dfa (так как эти нетерминалы раскрываются в символы алфавита dfa).

5 Асимптотика

Вообще получается, что на каждое правило p в грамматике мы добавляем где-то $O(|P|*|Q||P|+1)$ нетерминалов ($|p|$ - длина правой части). Но мы же знаем, что есть нормальная форма Хомского, у которой длина правых частей правил не больше двух. Вообще, нормальная форма Хомского нам не сильно нужна, нам достаточно привести грамматику к такому виду, чтобы длина правых частей все правил была меньше двух. Это делается довольно просто, так что в реализации алгоритма я полагаю, что это условие выполнено (хотя он и для произвольных грамматик работает, но только уж очень долго и уж очень большую грамматику строит).

Так что при ограничении на длину правил количество новых правил есть $O(|P| * |Q|^3 + |V_t|)$ ($|P|$ - кол-во правил в старой грамматике), а это и асимптотика по времени, и по памяти (длинна правил константна).

6 Конец

Спасибо за внимание)