

# Синтаксический анализ графов и задача генерации строк с ограничениями

Рустам Азимов, Семён Григорьев  
Лаборатория языковых инструментов JetBrains,  
Санкт-Петербургский государственный университет,  
Россия, 199034, Санкт-Петербург, Университетская наб. 7/9/  
rustam.azimov19021995@gmail.com, Semen.Grigorev@jetbrains.com

17 марта 2017 г.

## Аннотация

Одной из задач, изучаемых в теории формальных языков, является задача генерации строк, удовлетворяющих заданной системе правил. С другой стороны, существует задача синтаксического анализа графов, то есть задача поиска путей в графе, метки на ребрах которых образуют строку, принадлежащую заданному формальному языку. В данной работе будет показана связь между этими двумя задачами.

**Ключевые слова:** синтаксический анализ графов, генерация строк, формальные языки, конъюнктивные грамматики.

В таких областях, как графовые базы данных [3; 6], биоинформатика [1], возникают задачи поиска путей в графах, удовлетворяющих определенным ограничениям. В качестве таких ограничений естественно выбрать формальный язык  $L$  [2] и искать пути в графе, соответствующие строкам из языка  $L$ . Задачи поиска путей в графе, которые используют такие ограничения с формальными языками, называются задачами *синтаксического анализа графов*. Данная задача также возникает при статическом анализе динамически формируемого кода, например динамических SQL-запросов или генераторов Web-страниц. В данном случае графом является представление регулярной аппроксимации множества возможных значений динамически формируемых строк.

Кроме того, существует задача генерации строк, суть которой в построении строк, принадлежащих некоторому формальному языку. В работе [9] приведены формулировки задачи генерации строк с дополнительными ограничениями.

Некоторые вариации задач синтаксического анализа графов могут быть сведены к задаче генерации строк. Так, например, в большинстве задач синтаксического анализа графов недостаточно просто определить существование пути, соответствующего строке некоторого формального языка  $L$ , но

также требуется предъявить такой путь. Так как все пути в графе соответствуют строкам из некоторого регулярного языка  $R$ , то в данной задаче требуется найти путь, соответствующий строке из языка  $L \cap R$ . Эта задача может быть решена с помощью генератора строк рассматриваемого пересечения языков. В рамках данной работы была поставлена задача исследования связей между задачей генерации строк [9] и некоторыми типами задач синтаксического анализа графов [4; 5], использующие контекстно-свободные и конъюнктивные [8] языки.

Язык, который порождается графом  $G$  и выделенными в нем вершинами  $m, n$ , обозначим  $L(G, m, n)$ . А язык, порождаемый грамматикой  $C$ , со стартовым нетерминалом  $a$  обозначим  $L(C, a)$ .

В контексте задач синтаксического анализа графов бывает необходимо отвечать на различного рода вопросы, связанные с искомыми в графе путями. Тип вопросов, на которые отвечает задача принято называть *семантикой запроса*.

Использование *relational* семантики запроса означает, что для нетерминала  $a$  и графа  $G$  необходимо построить множество  $\{(m, n) \mid L(C, a) \cap L(G, m, n) \neq \emptyset\}$ . В случае использования КС-языка было выявлено отсутствие необходимости в применении генератора строк для поиска ответа на запрос с *relational* семантикой, так как в работе [5] используется аннотированная грамматика, которая порождает язык  $L(C, a) \cap L(G, m, n)$  и ее построение автоматически решает поставленную задачу.

Использование *all-path* семантики запроса означает, что для нетерминала  $a$ , графа  $G$  и его вершин  $m, n$ , необходимо предъявить все пути из вершины  $m$  в вершину  $n$ , такие что метки на ребрах этих путей образуют строку из языка  $L(C, a)$ . В случае использования КС-языка также было выявлено отсутствие необходимости в применении генератора строк для данной семантики, так как в работе [5] аннотированную грамматику и предлагают в качестве ответа на запрос. Но также была выявлена возможность использования генератора строк для получения конкретных строк пользователем из полученной аннотированной грамматики.

Использование *single-path* семантики запроса означает, что для нетерминала  $a$ , графа  $G$  и его вершин  $m, n$ , необходимо предъявить какой-нибудь путь (если он существует) из вершины  $m$  в вершину  $n$ , такой что метки на ребрах этого пути образуют строку из языка  $L(C, a)$ . Для КС-языков в работе [5] строится аннотированная грамматика, и если она порождает непустой язык, то в ней ищется строка минимальной длины, которая и будет соответствовать искомому пути в графе  $G$ . Таким образом, было выявлено, что алгоритм решения задачи синтаксического анализа графов с использованием *single-path* семантики запроса, предложенный в работе [5], и является примером использования генерации строки из КС-языка  $L(C, a) \cap L(G, m, n)$ .

Также была рассмотрена задача синтаксического анализа графов с использованием конъюнктивной грамматики. Из неразрешимости задачи определения пустоты конъюнктивных языков была получена неразрешимость задачи синтаксического анализа графов с использованием конъюнктивных языков и *relational* семантики запроса, о чем также упоминается в рабо-

те [4]. Кроме того, было выявлено, что при использовании конъюнктивных грамматик нельзя гарантировать нахождения хотя бы одной строки из конъюнктивного языка  $L(C, a) \cap L(G, m, n)$ . Предположим, что найдется хотя бы одна строка, удовлетворяющая рассматриваемым ограничениям. Тогда при использовании *all-path* семантики запроса, применяя алгоритм генерации строки, происходил бы просто перебор всех возможных строк и проверка на принадлежность этих строк к языку  $L(C, a) \cap L(G, m, n)$ , что не соответствует практическому смыслу задачи. А для задачи синтаксического анализа графов с использованием *single-path* семантики запроса есть возможность сгенерировать некоторую строку непустого языка  $L(C, a) \cap L(G, m, n)$ . Стоит отметить, что использование конъюнктивных языков в задачах синтаксического анализа графов мало изучено. Полученные результаты могут быть использованы в дальнейших исследованиях данной области. Одной из тем таких исследований, например, является применимость булевых [7] грамматик в синтаксическом анализе графов.

## Список литературы

1. *Anderson J., Novák Á., Sükösd Z.* Quantifying variances in comparative RNA secondary structure prediction // BMC Bioinformatics. — 2013. — С. 14–149. — ISBN 1471-2105-14-149.
2. *Barrett C., Jacob R., Marathe M.* Formal-language-constrained path problems // SIAM Journal on Computing. — 2000. — Т. 30, № 3. — С. 809–837.
3. Context-free path queries on RDF graphs / X. Zhang [и др.] // International Semantic Web Conference. — Springer. 2016. — С. 632–648.
4. *Hellings J.* Conjunctive context-free path queries.
5. *Hellings J.* Querying for Paths in Graphs using Context-Free Path Queries // arXiv preprint arXiv:1502.02242. — 2015.
6. *Mendelzon A., Wood P.* Finding Regular Simple Paths in Graph Databases // SIAM J. Computing. — 1995. — Т. 24, № 6. — С. 1235–1258.
7. *Okhotin A.* Boolean grammars // Information and Computation. — 2004. — Т. 194, № 1. — С. 19–48.
8. *Okhotin A.* Conjunctive grammars // Journal of Automata, Languages and Combinatorics. — 2001. — Т. 6, № 4. — С. 519–535.
9. *Охотин А.* О сложности задачи генерации строк // Дискретная математика. — 2003. — Т. 15, № 4. — С. 84–99.