Санкт-Петербургский государственный университет

Кафедра Системного программирования

Ершов Кирилл Максимович

Синтаксический анализ графов с помеченными вершинами и ребрами

Курсовая работа

Научный руководитель:
ст. преп., к. ф.-м. н. Григорьев С. В.

Санкт-Петербург
2016

Оглавление

Введение

Помеченные графы являются удобным способом представления различных структурированных данных. Такие графы используются, например, в биоинформатике, логистике, графовых базах данных.

Иногда для представления данных с использованием графов обходятся только метками на рёбрах. Но в некоторых случаях метки на вершинах позволяют более наглядно отображать зависимости между сущностями. К примеру, в биоинформатике существует большое количество данных, содержащих взаимосвязь между генами и белками. Такие данные удобно представлять в виде графа, вершины которого помечены определенными генами и белками, а ребра показывают их отношение (например, ген кодирует белок).

Часто возникает необходимость извлекать из графа пути, удовлетворяющие какому-либо запросу. Одним из способов задавать классы путей являются КС-грамматики. Пути рассматриваются как строки, состоящие из меток на рёбрах и вершинах. Остаётся проверить, принадлежит ли эта строка данному КС-языку. Для всех КС-грамматик существует эффективный алгоритм синтаксического GLL [3], основанный на идее рекурсивного спуска. На основе этого алгоритма и планируется реализовать возможность выполнять запросы к помеченным графам.

1. Постановка задачи

- В рамках проекта YaccConstructor [5] реализовать алгоритм на основе GLL, выполняющий поиск путей в графе с помеченными вершинами и рёбрами по заданной КС-грамматике
- протестировать алгоритм на реальных данных и сравнить производительность с существующими решениями

2. Обзор

Для реализации запросов к помеченным графам широко используются регулярные грамматики. Однако их возможностей бывает недостаточно для формулирования нужного запроса. Поэтому хотелось бы иметь возможность писать более выразительные запросы к графам, используя КС-грамматики.

2.1. Синтаксический анализ КС-грамматик

Для синтаксического анализа строки по произвольной КС-грамматике существуют различные алгоритмы. Например, Early parser [1] осуществляет разбор входной последовательности сверху-вниз и использует принцип динамического программирования. Этот алгоритм для произвольных КС-грамматик работает за время $O(n^3)$, для однозначных за $O(n^2)$ и за линейное время для большинства LR(k) грамматик. Динамический алгоритм CYK [6] также работает за время $O(n^3)$, однако требует приведения КС-грамматики к нормальной форме Хомского. Алгоритмы GLR [7] и GLL [3] являются обобщенными версиями анализаторов LR и LL соответственно (поддерживают неоднозначные грамматики), позволяют использовать произвольные КС-грамматики и работают в худшем случае за время $O(n^3)$. В основе GLL лежит нисходящий анализ, а значит он более прост в реализации. Алгоритм GLL для LL грамматик работает за линейное время.

2.2. Синтаксический анализ графов

Нахождение путей в графе с помеченными вершинами и рёбрами по КС-грамматике можно свести к задаче поиска путей в графе без помеченных вершин. Это можно сделать просто заменив вершины графа на ребро и две вершины, расположив метку с вершины на новом ребре. Это приведёт к увеличению числа вершин графа в 2 раза, а ребер прибавится на число вершин исходного графа. При достаточно больших входных данных это плохо скажется на производительности. Например, в работе

[4] была поставлена задача поиска связного подграфа в графе с помеченными вершинами и рёбрами по заданной КС-грамматике. Для проверки принадлежности пути КС-языку использовался алгоритм Early, а граф предварительно сводился к графу с метками только на рёбрах. Алгоритм тестировался на реальных данных, и для 300 пар вершин с максимальной длиной пути 8 время работы достигало 240 секунд, что делает этот алгоритм мало применимым на практике.

На кафедре СП Артёмом Гороховым в YaccConstructor был реализован алгоритм [8] на основе GLL, выполняющий запросы к графам с помеченными рёбрами по конъюктивной грамматике. Такие грамматики расширяют класс КС-грамматик. При описании продукций грамматики используется операция конъюнкции, что даёт возможность отсеивать ненужные цепочки. Однако тесты показали, что с конъюктивными грамматиками алгоритм работает в несколько раз медленнее, чем с контекстно-свободными.

2.3. YaccConstructor

На кафедре Системного программирования в лаборатории языковых инструментов разрабатывается проект YaccConstructor. Это платформа для исследований в области синтаксического анализа, написанная на языке F#. YaccConstructor позволяет создавать синтаксические анализаторы и имеет модульную архитектуру. Для построения анализатора выбирается фронтенд для обработки грамматик, выполняются необходимые преобразования и по указанному генератору строится нужный результат. В рамках этого проекта реализованы различные варианты алгоритма GLL.

3. Заключение

Результаты, достигнутые на данный момент:

- Написан обзор предметной области
- Реализован прототип, который находит начало и конец пути в графе с помеченными вершинами и рёбрами по заданной КС-грамматике

В дальнейшем планируется добавить в алгоритм поддержку SPPF (shared packed parse forest) [2] для вывода найденных путей, а также протестировать алгоритм на реальных данных и сравнить с существующими решениями.

Список литературы

[1] Earley Jay. An Efficient Context-Free Parsing Algorithm // Communications of the ACM. — 1970. — Vol. 13, no. 2. — P. 94–102.

[2] Rekers Joan Gerard. Parser generation for interactive environments : Ph. D. thesis / Joan Gerard Rekers ; Citeseer. — 1992.

[3] Scott Elizabeth, Johnstone Adrian. GLL parsing // Electronic Notes in Theoretical Computer Science. — 2010. — Vol. 253, no. 7. — P. 177–189.

[4] Sevon Petteri, Eronen Lauri. Subgraph queries by context-free grammars // Journal of Integrative Bioinformatics. — 2008. — Vol. 5, no. 2. — P. 100.

[5] YaccConstructor. YaccConstructor // YaccConstructor official page. — URL: `http://yaccconstructor.github.io`.

[6] Younger Daniel H. Recognition and parsing of context-free languages in time n^3 // Information and Control. — 1967. — Vol. 10. — P. 189–208.

[7] Younger Daniel H. LR parsers for natural languages // 10th International Conference on Computational Linguistics. — 1984. — Vol. 10. — P. 354–357.

[8] Артем Горохов. Поддержка конъюнктивных грамматик в GLL.