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
2017

Оглавление

Введение

1. Постановка задачи

- В рамках проекта YaccConstructor [2] реализовать возможность поиска путей в графе с помеченными вершинами и рёбрами по заданной КС-грамматике.
- Реализовать удобный интерфейс для работы:
  - создание и выполнение запросов
  - получение и обработка результатов
- Провести апробацию и сравнить с существующими решениями.

2. Обзор

2.1. Синтаксический анализ графов

2.2. YaccConstructor и QuickGraph.Query

3. Реализация

В YaccConstructor есть абстрактная реализация алгоритма синтаксического анализа GLL. Исходная грамматика описывается на языке спецификации грамматик YARD. Затем генератором из неё извлекается необходимая для работы алгоритма информация о грамматике. Во время выполнения алгоритм перемещается по входному объекту в зависимости от текущей позиции в грамматике. Объект, в котором требуется найти пути, удовлетворяющие исходной КС-грамматике, должен реализовывать интерфейс IParserInput. В рамках данной работы был реализован этот интерфейс для графов с помеченными вершинами и рёбрами. Если текущая позиция — вершина, следующими позициями в графе являются все исходящие рёбра. Если текущая позиция на ребре, следующей является конечная вершина. Таким образом, алгоритмом проверяются все возможные пути в графе. Результатом работы программы является sppf (лес разбора). Также для графа можно задать вершины, с которых будет начинать работу алгоритм и вершины, являющиеся конечными для синтаксического анализа. Для проверки работы алгоритма были написаны тесты.

В проекте QuickGraph.Query есть метод, извлекающий подграф из sppf. Но возвращает он граф с метками только рёбрах. Дополнительно была реализована возможность извлечения подграфа с метками на вершинах и рёбрах. Также реализована печать графа с метками на вершинах в dot-файл. Этот формат удобен для графического представления графов.

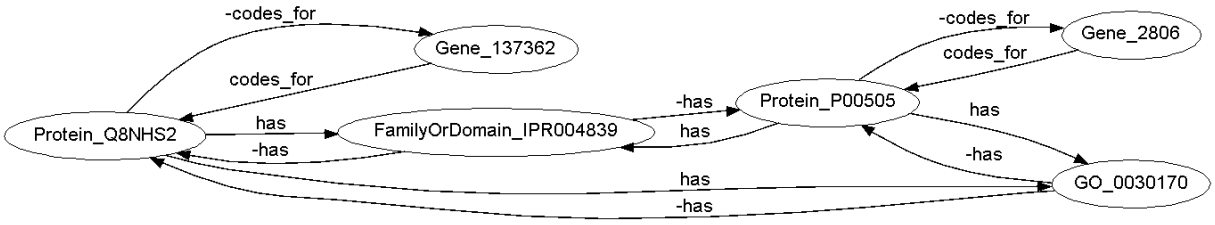

Рис. 1: Пример подграфа

4. Эксперименты

4.1. Данные

Существует большое количество биологических баз данных с открытым доступом, информация в которых может быть представлена как помеченный граф, в котором вершины соответствуют сущностям (протеины, гены, фенотипы), а рёбра отношениям между ними (взаимодействует, кодирует). Пути между вершинами позволяют найти новые связи в данных, либо показывают уже известные отношения. Подграф, построенный на всех найденных путях, более наглядно демонстрирует связи между вершинами. На рисунке 1 показан пример подграфа, построенного на путях между генами.

Реальный набор биологических данных был собран из разных баз данных, находящихся в открытом доступе: Entrez Gene (информация о генах), UniProt (протеины), Gene Ontology (биологические процессы), STRING (связи между протеинами), InterPro (семейства белков), KEGG (связи между генами), HomoloGene (группы гомологий генов). Данные были ограничены набором из пяти организмов: Homo sapiens, Rattus norvegicus, Mus musculus, D. melanogaster и C. elegans. Объединенные в один файл данные состоят из троек: субъект, отношение, объект. Такие тройки образуют помеченный ориентированный граф.

4.2. Запросы

Все вершины в полученном графе имеют уникальную метку. Но для удобства будем различать их по типу: гены, фенотипы и т.д. Назовём

```
[<Start>]
         s : gene
         v : protein | gene | GO | PATHWAY | FAMDOM | HOMOLOGENE
   similar : CODESFOR v RCODESFOR | BELONGS v RBELONGS
             | HAS v RHAS | HOMOLOGTO v RHOMOLOGTO
        ps :  (PROTEIN similar) *[1..2]
   protein : ps PROTEIN | PROTEIN
        gs : (GENE similar) *[1..2]
      gene : gs GENE | GENE
```

Рис. 2: Грамматика на языке YARD

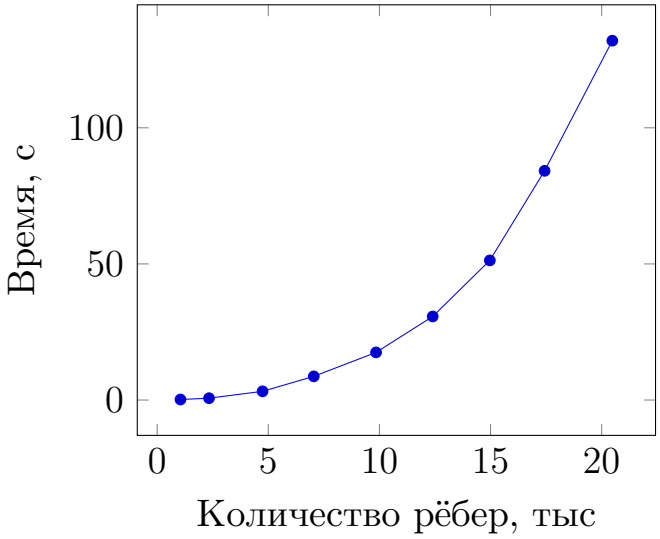

Рис. 3: Время работы алгоритма

две вершины в графе похожими, если они одного типа и имеют рёбра одного типа к похожим вершинам. Это определение рекурсивно. Таким образом, путь между похожими вершинами представляет собой палиндром, который нельзя задать с помощью регулярной грамматики. На рисунке 2 показана КС-грамматика на языке YARD, которая определяет похожие гены.

4.3. Производительность

Для оценки производительности была проведена серия экспериментов. Результаты приведены на графике, изображённом на рисунке 3. В статье [1] был проведён похожий эксперимент, но длины путей были ограничены от 4 до 8. В данной работе добиться такого ограничения не удалось, подграф строится по путям любой длины, поэтому нет возможности напрямую сравнить результаты.

5. Заключение

Список литературы

[1] Sevon Petteri, Eronen Lauri. Subgraph queries by context-free grammars // Journal of Integrative Bioinformatics (JIB). — 2008. — Vol. 5, no. 2. — P. 157–172.

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