Синтаксический анализ графов с помеченными вершинами и рёбрами

Ершов Кирилл Максимович, 344 группа
Научный руководитель:
ст. пр., к.ф-м.н. Григорьев Семён Вячеславович

Введение

- Помеченный граф — естественное представление различных структурированных данных:
  - биоинформатика
    - гены, белки, фенотип, их взаимосвязь
  - графовые БД
  - логистика
- Для эффективной работы с помеченными графами необходимо иметь возможность выполнять запросы

Введение

- Запросы можно представлять в виде грамматик
- Регулярные языки запросов:
  - Cypher Query Language
  - SPARQL
- КС-грамматики позволяют описывать более выразительные запросы
- Алгоритм синтаксического анализа GLL
  - $O(n^3)$
  - Реализован в YaccConstructor
- Подграф демонстрирует связи между вершинами более наглядно

Постановка задачи

- В рамках проекта YaccConstructor реализовать возможность поиска путей в графе с помеченными вершинами и рёбрами по заданной КС-грамматике
- Реализовать возможность извлечения подграфа, состоящего из путей, которые являются результатом КС-запроса к графу с метками на вершинах и рёбрах
- Провести экспериментальное исследование для реализованного алгоритма

Существующие решения

- Subgraph Queries by Context-free Grammars, 2008, Petteri Sevon and Lauri Eronen
  - Извлечение связного подграфа
  - Алгоритм Earley
  - Граф с метками на рёбрах и вершинах преобразуется в двудольный граф с метками только на рёбрах
- Context-Free Path Queries on RDF Graphs
  - RDF — модель представления данных в виде набора утверждений
  - Поиск путей в данных, представленных в формате RDF
  - Уступает по времени работы алгоритму GLL из YaccConstructor более чем в 10 раз

Существующие решения

- Conjunctive Context-Free Path Queries
  - Выполнение КС-запросов к графу
  - Основан на CYK
  - Реализован для графов с метками только на рёбрах

Выполнение запросов к графу

- Алгоритм GLL для графов в YaccConstructor
  - Все позиции в графе должны иметь уникальный номер
    - Вершины нумеруются чётными номерами
    - Все исходящие рёбра из вершины с номером k имеют позицию k+1
    - Нумерация осуществляется при добавлении рёбер к графу
  - Перемещение по графу в зависимости от текущей позиции в грамматике
    - По текущей позиции k в графе необходимо получать следующий токен и следующую позицию
    - Для эффективного получения всех исходящих рёбер используется структура AdjacencyGraph из библиотеки QuickGraph
  - Возможность задавать стартовые и конечные вершины

Подграф

- Извлечение подграфа
  - В результате выполнения запроса к графу получается SPPF
  - В разрабатываемом расширении библиотеки QuickGraph реализована возможность извлечения подграфа с метками на рёбрах из SPPF, где в вершинах указаны позиции во входном графе
  - Для получения подграфа с метками на вершинах и рёбрах, для каждой вершины с чётной позицией извлекаются метки с трёх следующих рёбер, которые соответствуют метке на начальной вершине, метке на ребре и метке на конечной вершине.

Эксперименты

- Данные
- Запросы
- Оценка производительности

Данные для апробации

- Биологические данные
  - Uniprot (протеины)
  - Entrez Gene (гены)
  - Gene Ontology (биологические процессы)
  - STRING (связи между протеинами)
  - KEGG (связи между генами)
- Организмы
  - Homo sapiens
  - Rattus norvegicus
  - Mus musculus
  - D. melanogaster
  - C. elegans

Запросы

- Поиск похожих вершин
- Каждая вершина имеет тип
- Назовём две вершины в графе похожими, если они одного типа и имеют рёбра одного типа к похожим вершинам. Это определение рекурсивно.
- Получились палиндромы, которые нельзя задать регулярной грамматикой

Запросы

```
[<Start>]
        s : gene
        v : protein | gene | GO | PATHWAY | FAMDOM
          | HOMOLOGENE
  similar : CODESFOR v RCODESFOR | BELONGS v RBELONGS
          | HAS v RHAS | HOMOLOGTO v RHOMOLOGTO
  protein : protein similar PROTEIN | PROTEIN
     gene : gene similar GENE | GENE
```

Рис.1: Грамматика на языке YARD, задающая похожие гены

Запросы

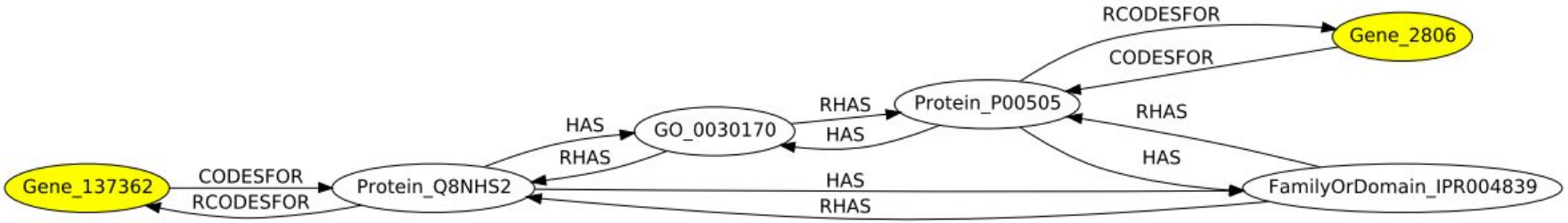


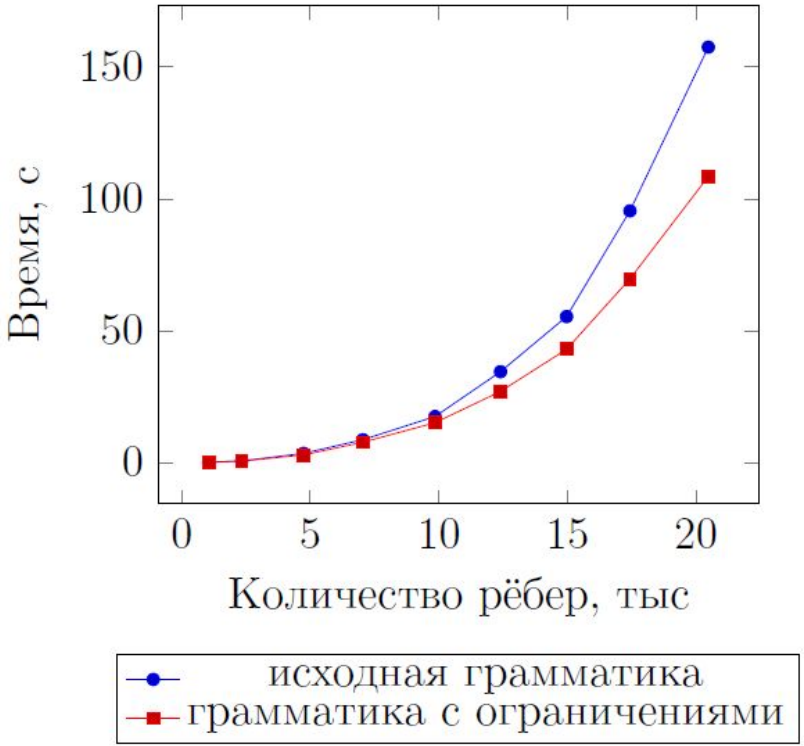
Рис. 2: Подграф, состоящий из путей между похожими генами

Запросы

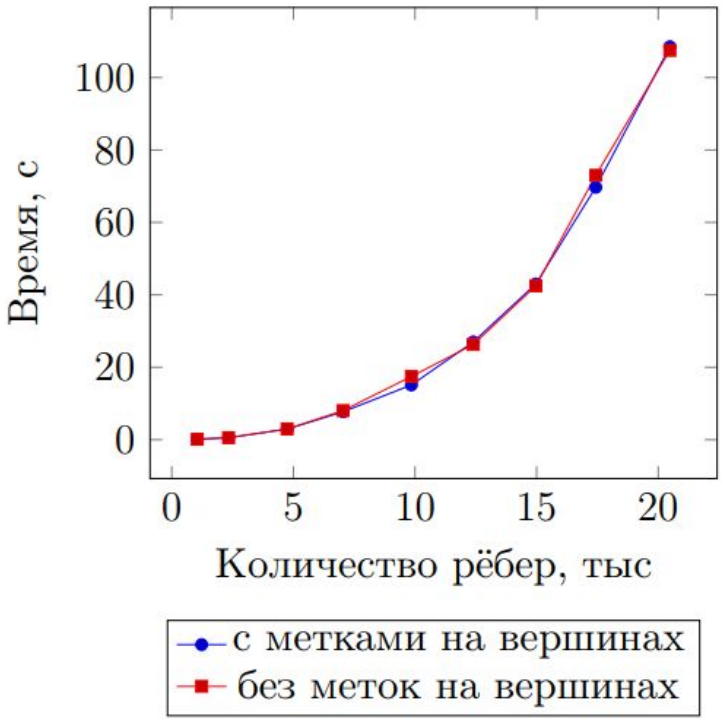
```
[<Start>]
        s : gene
        v : protein | gene | GO | PATHWAY | FAMDOM
            | HOMOLOGENE
  similar : CODESFOR v RCODESFOR | BELONGS v RBELONGS
            | HAS v RHAS | HOMOLOGTO v RHOMOLOGTO
       ps :  (PROTEIN similar) *[1..2]
  protein : ps PROTEIN | PROTEIN
       gs : (GENE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

Рис.1: Грамматика с ограничениями, задающая похожие гены

Оценка производительности



Оценка производительности



Результаты

- Реализована возможность поиска путей в графе с помеченными вершинами и рёбрами по заданной КС-грамматике
- Реализована возможность извлечения подграфа, состоящего из путей, которые являются результатом КС-запроса к графу с метками на вершинах и рёбрах
- Проведено экспериментальное исследование для оценки производительности