Super Duper Dataset for Experimental Analysis of CFPQ Algorithms

Author 1*
Author 2*
author_1@samplemail.com
author_2@samplemail.com
institution
city, state, country

Author 3
author_3@samplemail.com
institution
city, state, country

АННОТАЦИЯ

В последнее время наблюдается рост интереса к решению задач, связанных с контекстно-свободными запросами на графах. Критически важной становится потребность в измерении производительности алгоритмов решающих подобные задачи. Тем не менее разработка наборов контрольных данных и стандартизированные процедуры оценки отстают, что препятствует продвижению вперед в этой области. Чтобы решить эти проблемы мы представляем коллекцию CFPQ_Data, в которой собраны наиболее популярные графы для проведения экспериментального анализа алгоритмов решающих задачи контекстно-свободных запросов на графах. Коллекция состоит из более чем 40 графов разного размера. Также мы предоставляем загрузчики данных и реализации наиболее популярных алгоритмов на основе Python, а также стандарт проведения экспериментов и базовых показателей работы алгоритма. Здесь мы даем обзор собранной коллекции, стандартизированных процедур исследования и проводим базовые эксперименты. Все наборы данных доступны на сайте https://github.com/JetBrains-Research/CFPQ_Data. Проведенные эксперименты полностью воспроизводимы из кода, доступного на сайте https://github.com/JetBrains-Research/CFPQ_PyAlgo.

*Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Author 1, Author 2, and Author 3. 2018. Super Duper Dataset for Experimental Analysis of CFPQ Algorithms. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Здесь надо написать про то, что такое CFPQ и зачем оно надо. А также про то, почему надо собрать коллекцию графов в одном месте Представление данных с помощью помеченных графов находит своё применение в биоинформатике [12], в статическом анализе кода и многих других областях. Всё более популярными становятся графовые базы данных [9]. При работе с такими данными зачастую возникают запросы навигации и поиска путей, удовлетворяющих заданным ограничениям. Результат обработки такого рода запросов, как правило, представляет собой набор отношений между вершинами графа. Один из естественных способов определить подобные отношения над помеченным графом — указать соответствующие пути, используя формальные грамматики над алфавитом меток рёбер. Такие отношения могут быть заданы с помощью регулярных или контекстно-свободных языков [17]. Поскольку путь в помеченном графе можно рассматривать как слово в формальном языке, то соответствующие запросы навигации естественным образом могут быть выражены с помощью контекстно-свободных грамматик. Таким образом встает вопрос о необходимости разработки и реализации алгоритмов поиска путей с контекстно - свободными ограничениям (далее CFPQ алгоритмы). Ввиду широкой применимости контекстно-свободных запросов в перечисленных выше практических областях, критически важной становится потребность в измерении производительности алгоритмов реализующих эти запросы. Для того, чтобы показать применимость алгоритма на практике, возникает необходимость проведения экспериментального анализа на данных, моделирующих реальные сценарии. Также это позволяет

исследователям сравнивать производительность предлагаемого ими решения с уже существующими. Однако поиск и подготовка необходимых для проведения экспериментального анализа данных могут занять весьма продолжительное время. Одним из решений таких проблем во многих областях исследований является использование единого стандартизированного набора данных. Например, в биоинформатике очень важно иметь набор данных для проверки производительности алгоритмов кластеризации [3] и проекции данных [14]. А в области машинного обучения необходимо иметь стандартный набор данных, позволяющий исследователям выбирать какой метод лучше подходит для решения конкретной задачи[10]. В области алгоритмов, реализующих контекстно-свободные запросы к помеченным графам, на данный момент в большинстве работ, даже недавних, эксперименты проводятся на фиксированном наборе мелкомасштабных, неразнообразных графов, с использованием нестандартизованных экспериментальных протоколов и базовых показателей, что затрудняет сравнение результатов из разных публикаций.

1.1 Present work

Здесь мы даем обзор CFPQ_Data. Коллекция состоит из более чем 40 графов из широкого диапозона областей. Все графы представлены в стандартном формате RDF на сайте коллекции. https://github.com/JetBrains-Research/CFPQ_Data. Для облегчения работы с данными мы предоставляем загрузчики данных и реализации наиболее популярных алгоритмов на основе Python, а также стандарт проведения экспериментов и базовых показателей работы алгоритма. Кроме того, мы сообщаем результаты экспериментального исследования, сравнивающего наиболее популярные алгоритмы в этой области.

1.2 Related work

Здесь пишем про уже имеющиеся работы в области. А также про представленные в работах графы. Так например про RDF надо написать, что-то вроде Small graphs is a set of popular semantic web ontologies. This set is introduced by Xiaowang Zhang in "Context-Free Path Queries on RDF Graphs". Потом еще про MemoryAliases что-то вроде MemoryAliases — real-world data for points-to analysis of C code. First part is a dataset form Graspan tool. The original data is placed here. This part is placed in Graspan folder. Second part is a part of dataset form "Demand-driven alias analysis for C". This part is placed in small folder. И не забыть про синтетические графы. WorstCase — graphs with two cylces; the query Brackets is a grammar for the language of correct bracket sequences. SparseGraph — graphs generated with NetworkX to emulate sparse data. The grammar provided is a variant of the same-generation query. ScaleFree — graphs generated by using the Barab'asi-Albert model of scale-free networks. Use with grammar ** an_bm_cm_dn**, which is a query for AnBmCmDn language. FullGraph — cycle graphs, all edges are labeled with the same token. Use with A_star queries, which produce full graph on that dataset. LUBM - the Lehigh University Benchmark graphs.

2 THE CFPQ_DATA COLLECTION

Краткое описание того, что собрано. Коллекция состоит из более чем 40 графов разного размера. Также мы предоставляем загрузчики данных и реализации наиболее популярных алгоритмов на основе Python, а также стандарт проведения экспериментов и базовых показателей работы алгоритма. **Возможно сказать про листинг с примером кода.** Например выложить на сайт питоновский ноутбук с примером применения. Подробное описание каждого графа и другую документацию вы сможете найти на сайте коллекции.

2.1 Graphs and grammars

Описываем коллекцию. Наша коллекция наборов данных охватывает графы из разных областей, предоставленные разными авторами. Здесь мы даем обзор некоторых репрезентативных областей и моделей графов.

RDF. Рассказать откуда взялись RDF. Small graphs is a set of popular semantic web ontologies. This set is introduced by Xiaowang Zhang in "Context-Free Path Queries on RDF Graphs".

MemoryAliases. Рассказать откуда взялись MemoryAliases. Потом еще про MemoryAliases что-то вроде MemoryAliases — real-world data for points-to analysis of C code. First part is a dataset form Graspan tool. The original data is placed here. This part is placed in Graspan folder. Second part is a part of dataset form "Demand-driven alias analysis for C". This part is placed in small folder.

LUBM. Рассказать почему отдельно выделили LUBM. LUBM - the Lehigh University Benchmark graphs.

Synthetic. Рассказать про важность синтетических графов. WorstCase — graphs with two cylces; the query Brackets is a grammar for the language of correct bracket sequences. SparseGraph — graphs generated with NetworkX to emulate sparse data. The grammar provided is a variant of the same-generation query. ScaleFree — graphs generated by using the Barab'asi-Albert model of scale-free networks. Use with grammar ** an_bm_cm_dn**, which is a query for AnBmCmDn language. FullGraph — cycle graphs, all edges

are labeled with the same token. Use with A_star queries, which produce full graph on that dataset.

2.2 Baselines methods (GraphBuilders, GraphLoaders GraphSavers)

Здесь описываем как реализовали загрузку и работу с графами и почему так. Любой граф их датасета можно либо построить (если он синтетический) либо загрузить с сайта коллекции (или по пути к имеющемуся в системе графу). А с помощью GraphSaver'а можно любой граф сохранить в нужном исследователю виде. **Возможно стоит упомянуть про грамматики.**

2.3 Evaluation methods (Evaluators)

Тут надо описать каким образом происходит запуск алгоритма на графе и грамматике из коллекции. Возможно описать как это выглядит технически. А также как можно написать свой алгоритм в имеющейся инфраструктуре.

3 EXPERIMENTAL EVALUATION

Наша цель здесь это представить некоторый стандарт того, как можно проводить эксперименты. Стоит описать, стандартный путь проведения эксперимента и классические базовые показатели извлекаемые из работы алгоритма.

3.1 Datasets

Здесь нужно описать какие графы мы взяли для проведения СВОЕГО эксперимента. Мы взяли RDF потому что это классика. Взяли MemoryAliases потому что они большие. Взяли WorstCase потому что это важный случай в теории.

3.2 Algorithms

Здесь нужно описать какие мы алгоритмы использовали в СВОЕМ эксперименте. Думаю достаточно взять пару матричных алгоритмов и тензорный.

3.3 Results and discussion

Здесь мы суммируем полученные результаты. Както описываем полученные результаты. Возможно сравниваем их с полученными в других статьях. Можно написать насколько это было просто провести эксперимент в имеющейся инфраструктуре. Если код эксперимента маленький, то можно его вставить.

Таблица 1: Таблица с результатами эксперимента

Graph	Grammar	Algorithm	Time
RDF	G_1	Tensor	0.001
RDF	G_1	Tensor	0.001
RDF	G_1	Tensor	0.001
RDF	G_1	Tensor	0.001

4 CONCLUSION

Описываем то, к чему пришла статья. Собрали графы, грамматики, сделали удобную инфраструктуру для проведения экспериментов. Улучшили сопоставляемость результатов экспериментов. Мы сделали обзор коллекции CFPQ_Data и сообщили о результатах экспериментального исследования, сравнивающего наиболее популярные алгоритмы на подмножестве данных. Мы считаем, что наша коллекция наборов данных будет способствовать дальнейшему прогрессу в обучении представлению графов, и что наши унифицированные процедуры оценки улучшат сопоставимость результатов. Мы с нетерпением ждем добавления новых наборов данных и рады вкладу сообщества, исследователей и практиков из других областей. Дальнейшая работа включает более обширное сравнение существующих алгоритмов и добавление новых графов.

5 ACKNOWLEDGMENTS

Мы очень благодарны всем, кто принимал участие в этой нелегкой работе.