Московский государственный университет имени М.В. Ломоносова
Механико-математический факультет
Кафедра вычислительной математики

Курсовая работа

Поиск упоминаний персон и научных тематик в новостях для выявления возможного конфликта интересов при экспертизе.

Студент: Гвоздев Михаил Александрович
Преподаватель: С.н.с Кривчиков Максим Александрович

Группа: 510

Москва
2022

Содержание

1. Введение

Введение — общее описание проблемной области информационного поиска и поиска имен в частности и формулировка задачи.

В наше время интернет является огромным хранилищем различной информации. Не вся она полезна для конкретных задач. Поэтому люди придумали поисковые системы. Первым этапом в любом поиске является поверхностный сбор информации. Его осуществляют при помощи поисковых пауков [2].

Поисковый паук это программа, которая автоматически обходит структурированный набор адресов (URL) и заносит полученные страницы в специальную базу данных. Из полученных текстов, в свою очередь, извлекаются новые URL и добавляются в конец исходного набора. Пауки различаются по типам [7], так существуют *периодические* пауки широкого назначения [6]. Они скачивают все указанные в списке URL, пока не наберут желаемое количество скаченных страниц и останавливаются. Эта процедура повторяется периодически, когда возникает необходимость в обновлении данных. Они не самые эффективные с точки зрения скорости исполнения, зато отсутствует возможность не обновления заранее заданной страницы. Также существуют *пошаговые* пауки, продолжающие свою работу непрерывно. Их задача не заканчивать обходить страницы и заменять наименее полезные страницы на более полезные по некоторому правилу [6]. В этом случае могут возникать различные осложнения из-за неясности как универсально определить полезность некоторой страницы. Если считать страницу важной по непрерывному количеству посещений ее, то можно удалить важную страницу на которой раз в месяц оплачиваются счета за важные услуги (например коммунальные). Или же считать полезными только те страницы на которые больше всего ежемесячная аудитория, что делает невозможным использование такого паука с некоторой узкой целью (поиска информации по своей узкой специальности). Следующим типом являются *распределенные* пауки[4], состоящие из многих пауков широкого назначения, которые проверяют URL только в определенной области интернета ограниченной географически. При этом есть единый сервер, контролирующий и распределяющий URL между ними. Таким образом достигается большая отказоустойчивость всей системы, хоть и существует некоторое ограничение скорости в силу использования пауков широкого назначения. Эту проблему призваны были решать *параллельные* пауки [1]. В отличие от *распределенных* они обрабатывают единый массив URL, которые разделены на несколько машин, обрабатывающих эти URL параллельно, а не последовательно.

Такое улучшение позволило повысить скорость выгрузки страниц. Подтипом пауков широкого назначения являются *фокусированные*. Они обходят и добавляют в список дальнейшего обхода только те URL, страница которых соответствует некоторой теме (допустим теме поискового запроса), используя при этом различные методы подсчета обратных ссылок. В Google Inc. в 1998 году использовали для этой цели PageRank [5]. Для определения формулы дадим следующие условия, пусть на странице A цитируются страницы $S_1, ..., S_n$ (присутствуют их URL), $d \in (0, 1)$ параметр затухания (в работе брали 0.85), $C(A)$ - общее количество цитируемых страниц на странице A. Тогда $PR(A) = (1-d) + d * (\frac{PR(S_1)}{C(S_1)} + \frac{PR(S_2)}{C(S_2)} + ... + \frac{PR(S_n)}{C(S_n)})$, где PR это PageRank. Существует множество вариантов, выбора параметров для построения фокусированных пауков. Поэтому они, в свою очередь, подразделяются на различные виды в зависимости от способа определения соответствия страницы теме и способа обработки страниц [3].

Дальше по абзацу на следующее:

1. Место хранения всех этих обкаченных данных (Разные виды бд реляционные не реляционные язык SQL)
2. Способы поиска в тексте конкретных слов (полнотекстовый)
3. Способы определения связности найденных слов (через графы знаний, мл, еще сказать про что вообще значит связность)

2. Способы поиска нужных новостных публикаций

В принципе описание работы CommonCrawl, почему используем его, что именно с него берем, как выбрали что брать.

3. Способы хранения нужных данных локально

Краткое повторение про виды бд и подвод к тому почему сделали то что сделали. Описание работы загрузчика, описание структуры sqlite бд.

3.1. Определения

Вот не знаю что тут определять

4. Способы поиска людей в новостных записях.

Краткое повторение как в принципе их в тексте ищут. Описание работы нашего поиска людей (fts5) и то как мы определяем что они связаны (python).

4.1. Определения

Тоже не знаю что тут определить. Разве что стемминг (мб его альтернативу лемматизацию которую мы не использовали) и полнотекстовый поиск.

5. Заключение

Краткий пересказ того что было сделано. Описание дальнейшей работы при помощи графов связей, возможно с использованием онтологий.

Либо можно попытаться запихнуть онтологии в эту работу в раздел о способах поиска людей в новостных изданиях. Сделать сравнение что было до, чего добились с ними. Причину добавления вижу в том, чтобы появились формулы/определения/леммы/теоремы в курсовой. Просто неуверен что это правильно на 5 курсе быть без теорем в курсовой. С другой стороны может ничего не выйти и оказаться что в онтологиях никакого мат результата быстро не сделать. Я недостаточно глубоко в них погрузился.

Список литературы

1. Parallel Crawlers [Электронный ресурс]. URL: `https://ra.ethz.ch/CDstore/www2002/refereed/108/index.html` (дата обращения: 05.05.2022).
2. AbuKausar Md., S. Dhaka V., Kumar Singh S. Web Crawler: A Review // International Journal of Computer Applications. 2013. № 2 (63). С. 31–36.
3. Batsakis S., Petrakis E. G. M., Milios E. Improving the performance of focused web crawlers // Data & Knowledge Engineering. 2009. № 10 (68). С. 1001–1013.
4. Boldi P. [и др.]. UbiCrawler: a scalable fully distributed Web crawler // Software: Practice and Experience. 2004. № 8 (34). С. 711–726.
5. Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine // Computer Networks and ISDN Systems. 1998. № 1 (30). С. 107–117.
6. Cho J., Garcia-Molina H. The Evolution of the Web and Implications for an Incremental Crawler С. 18.
7. Wireless Communication and Computing) student, CSE Department, G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur, India [и

др.]. Study of Web Crawler and its Different Types // IOSR Journal of Computer Engineering. 2014. № 1 (16). С. 01–05.