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

В наше время интернет является огромным хранилищем различных данных. Не все они одинаково полезны для конкретных задач. В данной работе рассматривается задача поиска экспертов в новостных публикациях, с целью нахождения связности между ними. Связность при этом может быть любого вида. Здесь мы считаем что связь это наличие совместного упоминания в публикации двух и более экспертов. Но чтобы этого достичь нужно научиться правильно искать нужные данные. Поисковые системы реализуют механизм получения срезов данных. Первым этапом в любом поиске является поверхностный сбор информации. Его осуществляют при помощи поисковых пауков [7].

Поисковый паук это программа, которая автоматически обходит определенный заранее список адресов (URL) и заносит полученные страницы в специальную коллекцию. Из полученных страниц, в свою очередь, извлекаются новые URL и добавляются в конец исходного набора. Так они работали изначально, пока весь объем данных в интернете был сравнительно мал. Пауки различаются по типам [22]. В частности, существуют *периодические* пауки общего назначения [13]. Они скачивают все указанные в списке URL, пока не наберут требуемое автору паука количество скачанных 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 обновить заранее заданные страницы. Также существуют *пошаговые* пауки. Они имеют постоянный размер коллекции и продолжают свою работу непрерывно. Их задача беспрерывно работать и заменять наименее ”полезные” страницы на более ”полезные” по некоторому правилу [13]. Такой вид пауков появился в ответ на то, что данные в интернете постоянно меняются. Одни страницы появляются, другие наоборот исчезают. Не все они одинаково ”полезные” и в силу ограниченности доступной памяти некоторые удаляются. Построение универсальной метрики полезности является отдельной сложной задачей. 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[10]. Они состоят из многих пауков общего назначения, которые проверяют URL

только в определенной области интернета, частое использование которой характерно для ограниченной географически территории. Например для определенной страны характерно использование сайтов в основном на ее языке. При этом есть центральный сервер, контролирующий и распределяющий URL между ними. Таким образом, достигается большая отказоустойчивость всей системы, хоть и существует некоторое ограничение скорости в силу использования пауков общего назначения. Эту проблему призваны были решать *параллельные* пауки [2]. 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 общего назначения являются *фокусированные*. Они обходят и добавляют в список дальнейшего обхода только те URL, по которым находится документ соответствующий некоторой теме (допустим теме поискового запроса). При этом используются различные методы подсчета обратных ссылок. Добавление новых ссылок происходит до тех пор пока не наберется нужное количество страниц или не обойдется весь список URL. В Google Inc. в 1998 году для подсчета обратных ссылок использовали PageRank [11].

Для описания алгоритма PageRank зададим 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 [9].

Когда мы научились правильно искать данные, возникает следующая задача - надежное хранение и быстрый доступ к ним. Отвечая на такой запрос появился язык SQL[18] для работы с реляционными базами данных. Реляционная модель представляет собой набор двумерных таблиц. Каждая таблица состоит из строк - записей и столбцов - полей. Поля обязаны быть одного из допустимых типов. Таблицы могут быть связаны друг с другом при помощи различных ключей (ссылок). Изначально реляционные базы данных создавались для хранения на одной машине сравнительного небольшого объема данных. Когда объем их достаточно вырос старая парадигма баз данных перестала работать. CAP теорема [1] утверждает, что любая система общих данных может обладать наибольшее двумя свойствами из следующих: согласованность - существует только одна актуальная версия данных, доступность - данные доступны в любой момент времени, стабильность - устойчи-

вость к физическим нарушениям связности частей данных. В связи с этим стали появляться нереляционные базы данных и системы для управления ими. Они специализированы под определенные задачи и поэтому могут делать незначительные для их цели допущения в CAP теореме. В наше время можно выделить следующие типы нереляционных моделей данных[15]. Наиболее известной является модель *ключ-значение*. Принцип ее работы похож на идею хеш-таблицы. У каждой записи есть уникальный ключ, которому соответствует единственный хранящий запись сервер. Следующий тип это *документный*. Его идея состоит в том, что документы устроены гораздо сложнее, чем просто текстовые поля. Они могут содержать ссылки на другие документы, которые в свою очередь ссылаются на дополнительные и так далее. *Столбцовая* модель отличается от предыдущих тем, что хранит данны в виде столбцов, а не записей. Основополагающей системой для этой модели является Bigtable[12]. В ней данные хранятся в виде наборов столбцов одинакового типа. При этом таких наборов может быть максимум сотни, в отличие от реляционной модели в которой может быть неограниченное количество столбцов. Дальше идет *графовая* модель. Она строится на основе модели графа из теории графов. Ее преимущество состоит в том что она позволяет эффективно обрабатывать данные с большим числом связей между объектами разной природы.

Следующий этап это поиск упоминаний в самом тексте страницы. В отличие от поисковых пауков он проходит весь текст содержащийся на ней, пытаясь найти нужную последовательность символов. Это можно делать как каждый раз проходя один и тот же текст или же строить инвертированный индекс[23]. Он представляет собой структуру данных, в которой каждому слову сопоставляются места, где оно упоминается в различных текстах. Прежде чем строить подобный индекс, слова нужно привести в некоторую единую форму. Одним из таких решений является стемминг. Этот процесс оставляет от слова только его основу, не обязательно являющуюся корнем этого слова. Одной из его наиболее популярных реализаций является стеммер Портера[21], который в последствии перерос в отдельный проект Snowball[3]. В общем виде все алгоритмы стемминга можно разбить на 3 группы [17]: усеченные, статистические и смешанные. Примером усеченного является алгоритм Портера. Статистических - концепция n-gram. Идея ее заключается в том, что похожие слова имеют очень высокий процент совпадения. Алгоритм разбивает исходное слово на части длины n. Так для слова ”слон” 3-граммами будут последовательности: ”--с”, ”-сл”, ”сло”, ”лон”, ”он-”, ”н--”, где ”-” означает пробел. Достоинством этой модели является ее независимость от языка, а недостатком долгая работа и большие

затраты памяти. При помощи статистических методов оценивается есть ли заданное слово в том или ином документе. Другим вариантом нормализации слов является лемматизация. Она приводит слова в начальную форму (учил -> учить, домов -> дом). Такой вариант позволяет сохранить больше контекста при поиске определенный фраз. Однако есть более высокая вероятность ошибиться и в итоге только усложнить поиск[8]. Смешанная модель стемминга же сочетает в себе достоинства первых двух. Построенный при помощи этих методов инвертированный индекс позволяет значительно ускорить работу поисковых систем.

Дальше подходящие страницы следует отсортировать от наиболее вероятного к наименее (релевантность) на основе более продвинутых методов. Базово это можно выполнять при помощи различных вариаций алгоритма TF-IDF [20], [19]. Основная идея его заключается в том, что можно разбить формулу определения релевантности на 2 множителя. TF (term frequency) описывает как часто искомое слово встречается в тексте страницы (документе). Например $\frac{w}{W}$, где w - количество совпадающих слов, W - число всех слов в документе. IDF (inverse document frequency) же является величиной пропорциональной отношению релевантных документов ко всем документам коллекции. В качестве нее можно брать например $log(\frac{N}{n})$, где n - число релевантных документов, а N - число всех документов. Итоговая формула для определения релевантности получается перемножением TF$\times$IDF, что в примере равно $\frac{w}{W}log(\frac{N}{n})$.

Перейдем к самой задаче данной работы. Нужно научиться находить ученых в полученных текстах и определять связаны ли они. Под связностью двух людей (рецензента и рецензируемого) будем понимать наличие знакомства, дружбы, родства или других отношений, которые могут оказать влияние на вынесение рецензентом вердикта на научную работу рецензируемого. Одним из вариантов, которыми это можно сделать, являются графы знаний[16]. Знанием в них понимается некоторое утверждение которое известно. Они бывают разных видов[14]. Наиболее известными являются дополняющийся еженедельно Wikidata[5] с более чем 700 миллионами записей и 18 миллионами объектами, Freebase - не обновляющийся с 3 миллиардами записей о 50 миллионах объектов и DBpedia - дополняющийся несколько раз в год краудсорсинговый проект с более чем 400 миллионами записей о более чем 4 миллионах объектов. Основным отличием графов знаний от обычных распределенных баз данных является наличие возможности выносить некоторые логические суждения об объекте. Другим вариантом решения проблемы могла бы являться модель word2vec, которая сопоставляет контексту/слову наиболее вероятное следующее слово/контекст.

Она создает специальные числовые векторы - эмбединги, для каждого слова. Таким образом можно было бы сравнивать контексты при помощи косинусного расстояния $\frac{|a||b|}{(a,b)}$, где a, b векторы эмбедингов. И делать выводы о связности двух имен на основе того насколько мало это расстояние. Однако существенным недостатком ее является необходимость в огромных вычислительных мощностях и больших объемах памяти.

2. Методы сбора данных

Чтобы не писать своего собственного поискового паука в этой работе используется паук широкого назначения CommonCrawl[6]. Он обходит весь интернет который позволяет себя индексировать и скачивает их себе в архив. Потом загружает их в общий доступ. Он был выбран так как очень часто новостные страницы используют скриптовые языки на своих сайтах. А это значит что их очень сложно обходить и выгружать с них сами тексты новостей.

CommonCrawl позволяет искать по определенным доменам и URL. Так, вводя домен ”.ru” вам будут даны ссылки на скачивание всех сайтов у которых доменом является ”.ru”. Или ”`https://www.kp.ru/`” вы получите ссылки на архивы для скачивания всей информации находящейся на этом сайте. Также можно указывать метаинформацию (например искать только страницы у которых в HTML структуре указано ”rus”), тем самым сужая область поиска.

3. Структура хранения данных

Как было сказано в введении базы данных бывают разных типов. В нашей конкретной задаче было достаточно обычной реляционной базы данных. Так как объемы информации не настолько большие чтобы выделять под это отдельный сервер и частота запросов невелика, было решено остановиться на базе данных Sqlite.

Загрузка информации происходит в несколько этапов:

1. Собираем список нужных новостных изданий.
2. Запись списка URL в базу данных в таблицу с именем queue колонку url.
3. Дальше в несколько процессов запускается скрипт загрузчика.
4. Загрузчик берет каждую запись из очереди.

5. При помощи API CommonCrawl получает ссылки на архивы для скачивания.
6. Пакетами, которые позволяет хранящий ресурс, данные загружаются в таблицу content.
7. Текст отправляется в колонку raw_cont, а адрес страницы в колонку url

Структура таблиц:

```
TABLE man(id primary key, lastname, firstname, middlename, expert);
TABLE queue(
    id INTEGER PRIMARY KEY AUTOINCREMENT,
    url TEXT,
    load_started DEFAULT NULL,
    load_complete DEFAULT NULL
);
TABLE content(
    id INTEGER PRIMARY KEY AUTOINCREMENT,
    url TEXT,
    cont TEXT
);
```

Всего было выгружено 5Гб очищенных данных. Загрузка производилась в несколько процессов. Данные очищались на этапе получения от различных HTML тегов при помощи python библиотеки BeautifulSoup и проекта readability.

3.1. Математическая модель структуры данных

Основана на реляционной алгебре. Так основным элементом являются мультимножества.

1. Отношение - это таблица или ее подмножество.
2. Кортеж - строка в талице.
3. Атрибут - столбец в таблице.
4. Поле - ячейка таблицы определенного типа.
5. Тип ячейки в таблице - некоторое число $a \in N$, где N - натуральное число.

   Соответствие типов:

(a) 1 - целое число
(b) 2 - текст (последовательность символов)
(c) 3 - булевое значение (0 - Ложь, 1- Истина)
(d) 4 - время

6. Множество уникальных элементов $U(t_1, ..., t_k)$ - набор уникальных кортежей, где у каждого кортежа есть k полей соответствующим атрибутам $t_1, ..., t_k$.

7. Мультимножество ($M(U(t_1, ..., t_k))$) построенное на множестве уникальных элементов $U(t_1, ..., t_k)$ - это упорядоченная пара (U, φ) такая, что $\varphi : U \to N$, где N - множество натуральных чисел, отображение $\varphi : \forall x \in U \; \exists y \in N : \varphi(x) = y$. (дальше множество уникальных элементов будет опускаться)

В нашем случае имеется 3 мультимножества: $man(1, 2, 2, 2, 3)$, $queue(1, 2, 4, 4)$, $content(1, 2, 2)$

4. Поиск упоминаний персон

Как и было указано ранее весь поиск конкретной информации в тексте можно разбить на несколько групп в зависимости от того какой способ нормализации мы используем. Существуют разные способы нормализации текстов. На их основе существуют различные способы поиска - статистическая, усекающая и смешанная. Также можно использовать регулярные выражения.

4.1. Определения

1. Полнотекстовый поиск - поиск который ведется по всему документу или по существенной его части.
2. Стемминг - процесс редуцирования слова до его основы или корня.
3. Лемматизация - приведение словоформы к начальной форме.
4. Начальная форма слова - единственная форма для каждого слова. В русском языке для существительных это единственное число, именительный падеж, глаголов - инфинитив, прилагательные - мужской род, единственное число и так далее.
5. Нормализация слова - приведение слова к его начальной форме.

4.2. Описание модели

В данной работе используется стемминг и на нем строится усекающая модель нормализации текстов, так как нет доступа к нужными вычислительными ресурсами для применения остальных.

Также так как было выгружено около 100 000 различных записей их нужно правильно индексировать и искать по ним, так как пройти 100 000 последовательностей из более чем 10 000 символов на Python довольно долго. Поэтому используем python с подключенной к нему sqlite и прямые SQL запросы. Для полнотекстового поиска используем fts5[4], это специальный набор команд для работы с виртуальными таблицами в sqlite и поддерживающий полнотекстовый поиск. При помощи нее создается B-дерево всех документов. Так получается кратно ускорить работу полнотекстового поиска по базе данных. Для самого полнотекстового поиска используется отбрасывание окончаний у имени, фамилии и отчества в итоге оставшаяся основа слова при помощи регулярного выражения прогоняется по всей базе текстов. На втором этапе уже составляется таблица пересечений людей в одном тексте (то есть встретились 2 одинаковые фамилии). Дальше уже на оставшихся текстах происходит прогон на наличие в тексте также соответствующих имен и отчеств. Этот этап позволяет избавиться от таких совпадений как для фамилии Орехов и слова орехи, фамилии Толстой и слова толстой. Поиск осуществляется не по всему тексту, а только по его небольшому интервалу вокруг предполагаемо найденной фамилии. В работе используется интервал от 50 до 100 символов, так как существуют двойные фамилии а также длинные южные.

5. Заключение

В данной работе построена математическая модель графа связей сотрудников научных организаций. Все представленные в работе методы обработки текстов находятся на начальном уровне и требуют доработки и улучшения. В будущем планируется расширение вида информации о каждом конкретном сотруднике. Это планируется сделать добавлением организаций, в которых они работают, места рождения/проживания, места обучения, предыдущие места работы. Потенциальное развитие работы представляется в повышении скорости загрузчика и объема данных на основе которых строится индекс. Также видится возможность развития при помощи добавления триплетов и более длинных n-плетов. На основе которых при помощи онтологий или моделей машинного обучения будут строиться логические заключения.

Список литературы

1. CAP Twelve Years Later: How the «Rules» Have Changed // InfoQ [Электронный ресурс]. URL: https://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed/ (дата обращения: 17.05.2022).
2. Parallel Crawlers [Электронный ресурс]. URL: https://ra.ethz.ch/CDstore/www2002/refereed/108/index.html (дата обращения: 05.05.2022).
3. Snowball: A language for stemming algorithms [Электронный ресурс]. URL: http://snowball.tartarus.org/texts/introduction.html (дата обращения: 20.05.2022).
4. SQLite FTS5 Extension [Электронный ресурс]. URL: https://www.sqlite.org/fts5.html (дата обращения: 22.05.2022).
5. Wikidata [Электронный ресурс]. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (дата обращения: 21.05.2022).
6. So you’re ready to get started. – Common Crawl [Электронный ресурс]. URL: https://commoncrawl.org/the-data/get-started/ (дата обращения: 21.05.2022).
7. AbuKausar Md., S. Dhaka V., Kumar Singh S. Web Crawler: A Review // International Journal of Computer Applications. 2013. № 2 (63). С. 31–36.
8. Balakrishnan V., Ethel L.-Y. Stemming and Lemmatization: A Comparison of Retrieval Performances // Lecture Notes on Software Engineering. 2014. № 3 (2). С. 262–267.
9. Batsakis S., Petrakis E. G. M., Milios E. Improving the performance of focused web crawlers // Data & Knowledge Engineering. 2009. № 10 (68). С. 1001–1013.
10. Boldi P. [и др.]. UbiCrawler: a scalable fully distributed Web crawler // Software: Practice and Experience. 2004. № 8 (34). С. 711–726.
11. 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.
12. Chang F. [и др.]. Bigtable: A Distributed Storage System for Structured Data // ACM Transactions on Computer Systems. 2008. № 2 (26). С. 1–26.
13. Cho J., Garcia-Molina H. The Evolution of the Web and Implications for an Incremental Crawler С. 18.
14. Färber M., Rettinger A. Which Knowledge Graph Is Best for Me? // 2018.
15. Grolinger K. [и др.]. Data management in cloud environments: NoSQL and NewSQL data stores // Journal of Cloud Computing: Advances, Systems and Applications. 2013. № 1 (2). С. 22.
16. Hogan A. [и др.]. Knowledge Graphs // ACM Computing Surveys. 2022. № 4 (54). С. 1–37.
17. Jivani A. G. A Comparative Study of Stemming Algorithms 2011. (2). С. 9.

18. Melton J. SQL language summary // ACM Computing Surveys. 1996. № 1 (28). С. 141–143.
19. Ramos J. Using TF-IDF to Determine Word Relevance in Document Queries.
20. Salton G., Buckley C. Information processing & management // Term-weighting approaches in automatic text retrieval. 1988. № 5 (24). С. 513–523.
21. Willett P. The Porter stemming algorithm: then and now // Program. 2006. № 3 (40). С. 219–223.
22. 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.
23. Zobel J., Moffat A. Inverted files for text search engines // ACM Computing Surveys. 2006. № 2 (38). С. 6.