Санкт-Петербургский государственный университет
Программная инженерия

Ахмедов Гаджи Омар оглы

Модели обработки естественного языка в задачах информационного поиска

Отчёт по учебной (ознакомительной) практике
Научный руководитель: к.ф.-м.н., доцент кафедры СП Д.В.Луцив

Санкт-Петербург
2021

Содержание

Введение:

Яндекс знают все. Но мало кто знает его историю – в чем секрет успеха. В 1993 году была написана первая рабочая версия приложения для локального поиска (на жёстком диске компьютера), которое назвали «Yandex». Слово расшифровывалось как **y**et **an**other in**dex**er (с англ. — «ещё один индексатор», «очередной индексатор»).

В 1993—1994 годы программисты CompTek начали сотрудничать с лабораторией компьютерной лингвистики Института проблем передачи информации, которой заведовал академик РАН Юрий Апресян. Сегалович в качестве основного разработчика написал программу автоматического морфологического анализа, которая использовалась при поиске.[1][2][3] Результатом совместного труда программистов стал словарь с поиском, учитывавшим морфологию русского языка, ещё одним его достоинством было то, что он целиком грузился в оперативную память и быстро работал.

В 1994 году на основе разработанных технологий программисты CompTek создали «Библейский компьютерный справочник» — информационно-поисковую систему, работавшую с текстом Библии. Для преобразования в электронный вид почти половину книги пришлось набирать вручную. С 1995 года компания работала над проектом «Академического издания классиков на CD ROM», который предполагал выход полного электронного академического издания Александра Грибоедова и Александра Пушкина со словарем языка Грибоедова. К 1996 году был разработан алгоритм построения гипотез: если искомого слова не было в словаре, то поиск осуществлялся по наиболее похожим на него, и по ним уже строилась модель словоизменения.

Все более актуальным становится использование различных программ, осуществляющих поиск документов различных форматов, информации в СУБД (Система управления базами данных) и информационных системах, сообщений электронной почты и прочих данных, содержащихся как на жестком диске персонального компьютера или в локальной сети предприятия, так и в других источниках знаний.[4] Востребованность поисковых систем такого рода обусловлена непрерывным ростом объемов текстовой информации. Но несмотря на это, приоритетным направлением развития поисковых технологий (помимо Интернета) является корпоративный сектор. Наиболее важным параметром любой системы поиска является скорость ее работы. Это касается как индексации больших объемов данных, так и скорости поиска документов. Конечно, немаловажными факторам являются

возможности работы с различными источниками данных, списки поддерживаемых форматов файлов и дополнительный функционал (поддержка морфологии, синонимов, различных видов поиска).

Проблема организации в одну базу данных частично решается за счет DMS, CRM и специализированных СУБД. Но, чем больше предприятие (в качестве примера можно взять и СПбГУ, у которого большой набор учебно-методической и нормативной документации) и чем разнообразнее его виды деятельности, тем сложнее обрабатывать информацию из различных источников. Документы на диске, 1С, Oracle, архивы html-страниц, электронная корреспонденция и даже записи логов ICQ (ICQ — бесплатная кроссплатформенная система мгновенного обмена сообщениями, для мобильных и иных платформ с поддержкой голосовой и видеосвязи.) – в последнее время отнюдь немаловажный «информационный сектор», который можно смело подключать к основным хранилищам данных внутри любой крупной компании. На основе анализа многообразия этих источников поступления и хранения текстовых данных можно выделить две основные проблемы «информационного обеспечения». Это не структурированность информации и ее поиск. В принципе, эти проблемы взаимосвязаны. Так как, получив хорошую систему поиска информации по различным источникам, можно, тем самым, предельно систематизировать полученные результаты [8].

1 Цели и задачи

Целью работы является создание десктопного поисковика, ориентированный на конкретный вид документации (Например: Учебно-методическую документацию). Для достижения обозначенной цели были поставлены следующие задачи:

• Объединение несколько видов поиска;

• Возможность создания выбора элементов для индексирования;

• Индексация электронной почты, тэгов аудио, видео файлов и изображений;

• И выбрать максимально эффективный способ в рамках ограниченных вычислительных ресурсов.

2 Проблемы с поиском

Функции настольного поиска, встроенные в современные операционные системы, программы электронной почты и другие приложения, обладают гораздо меньшими возможностями, чем поисковые системы в Интернете. Как правило, они предлагают только простой поиск по ключевым словам в наборе файлов, как правило, одного типа.

В Интернете поисковые системы могут использовать информацию, организованную в общий формат HTML, со стандартными способами идентификации различных элементов документа. Механизмы могут использовать эту информацию вместе со ссылками на другие документы, чтобы делать статистические предположения, повышающие вероятность получения релевантных результатов. Поиск на рабочем столе более сложен, поскольку Microsoft Word и другие приложения по-разному форматируют документы разных типов. Кроме того, файлы рабочего стола могут быть как структурированными, так и неструктурированными.

Функция и значение структурированных файлов, таких как информация в реляционной базе данных или текстовый документ со встроенными тегами, четко отражены в их структуре. Легко идентифицируемая структура облегчает поиск таких файлов. Это не относится к неструктурированной информации, которая включает в себя документы на естественном языке, неформатированные текстовые файлы, речь, аудио, изображения и видео.[11][12] Таким образом, поисковые системы для настольных ПК должны добавлять возможности иначе, чем приложения веб-поиска.

Однако поисковые системы настольных компьютеров сталкиваются с дополнительной проблемой распознавания того, с каким из многих типов файлов они имеют дело. Механизмы также должны получать любые метаданные, которые авторы решили включить в заметки электронной почты, файлы базы данных и другие типы документов. При проведении поиска настольные движки должны быть эффективными и не создавать значительных вычислительных ресурсов или нагрузки на память компьютера.

Служба веб-поиска может выделить целую ферму серверов для выполнения только поиска, в то время как поисковая система для настольных компьютеров должна быть максимально эффективной в рамках ограничений вычислительных ресурсов пользователя.

2.1 Настольные поисковые системы.

Настольные поисковые системы используют одну или несколько программ-сканеров файлов, подобных тем, которые используются поисковыми системами в Интернете, которые после установки перемещаются по дискам. Поисковые роботы используют индексатор [13] для создания индекса файлов; их расположение в иерархической древовидной файловой структуре жесткого диска; имена файлов, типы и расширения (например, .doc или .jpg); и ключевые слова. Как только существующие файлы проиндексированы, сканер индексирует новые документы в режиме реального времени. Во время поиска механизм сопоставляет запросы с проиндексированными элементами, чтобы быстрее находить нужные файлы.

Поисковые роботы также собирают метаданные, которые позволяют движку более разумно обращаться к файлам, предоставляя дополнительные параметры поиска. Несколько настольных поисковых систем интегрированы с веб-системами провайдеров и одновременно выполняют оба типа поиска по запросам. Как по мне эти поставщики прилагают значительные усилия для создания наборов функций и интерфейсов для настольных компьютеров, которые будут такими же знакомыми и простыми в использовании, как и их веб-аналоги.[14]

2.2 Семантическое представление текста

Семантический анализ – важная подзадача обработки естественного языка (NLP (Natural language processing)), этап в последовательности действий алгоритма автоматического понимания текстов, заключающийся в выделении семантических отношений, формировании семантического представления текстов. В общем случае семантическое представление является графом, семантической сетью, отражающей бинарные отношения между двумя узлами — смысловыми единицами текста.

2.2.1 Bert

Bert — это нейронная сеть от Google, показавшая с большим отрывом state-of-the-art результаты на целом ряде задач.[15][16][17] С помощью BERT можно создавать программы с ИИ для обработки естественного языка: отвечать на вопросы, заданные в произвольной форме, создавать чат-ботов, автоматические переводчики, анализировать текст, а в моём случаи десктопный поисковик и так далее.

Существует несколько способов представлять слова векторами, они постепенно эволюционировали: word2vec, GloVe, Elmo. Хотелось бы по

подробнее остановиться на одном из них . Например: word2vec.

2.2.2 Введение в Word2vec

Word2vec — это продукт распределительной семантики, которая является одной из самых успешных идей современного статистического НЛП. Распределительная семантика — это область исследований, которая концентрируется на количественной оценке и категоризации семантических сходств между языковыми единицами на основе их свойств распределения. Word2vec как неконтролируемая техника полностью достигает цели.

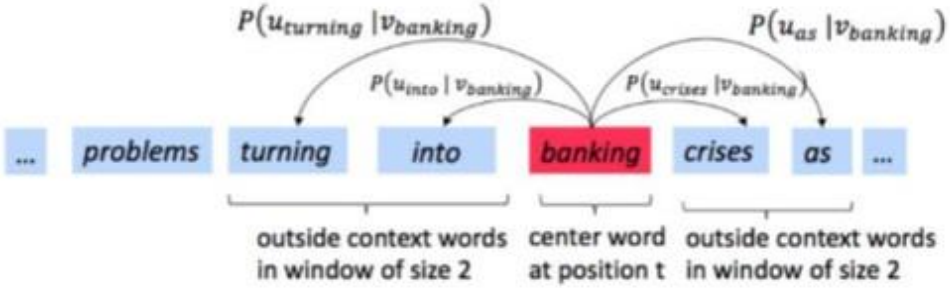

Рисунок 2.1 Предсказание слова

Как мы видим, условная вероятность является основой этой теории. Затем для каждой позиции t = 1,...,T, заданное центральное слово wt, предсказывает слова контекста в пределах окна фиксированного размера m, L(θ) = QT t=1Qm≥j≥−m p(wt+j|wt;θ). Теперь мы получаем целевую функцию J(θ) = −1 T log(L(θ))PT t=1Pm≥j≥−m p(wt+j|wt;θ). Наша цель состоит в том, чтобы минимизировать функцию объекта, которая в равной степени максимизирует точность прогнозирования. Мы используем функцию softmax для вычисления p(wt+j|wt;θ). Функция softmax pi = exp(xi) Pn j=1 exp(xj) отображает произвольные значения xi в распределение вероятностей pi. Достигнув производной целевой функции по вектору центрального слова vc, мы получаем ∂ ∂vc logp(o|c) = uo −Pv x=1 p(x|c)ux. Эта разница, оказывается, точно дает нам наклон, в котором мы должны идти и изменение представления слова, чтобы улучшить способность нашей модели прогнозировать.

2.2.3 Нейронная сеть Word2vec

Word2vec использует нейронную сеть для обучения. Есть один входной слой, в котором столько нейронов, сколько слов в словаре для обучения. Второй слой — это скрытый слой, последний слой — это выходной слой, который имеет то же количество нейронов, что и входной слой.

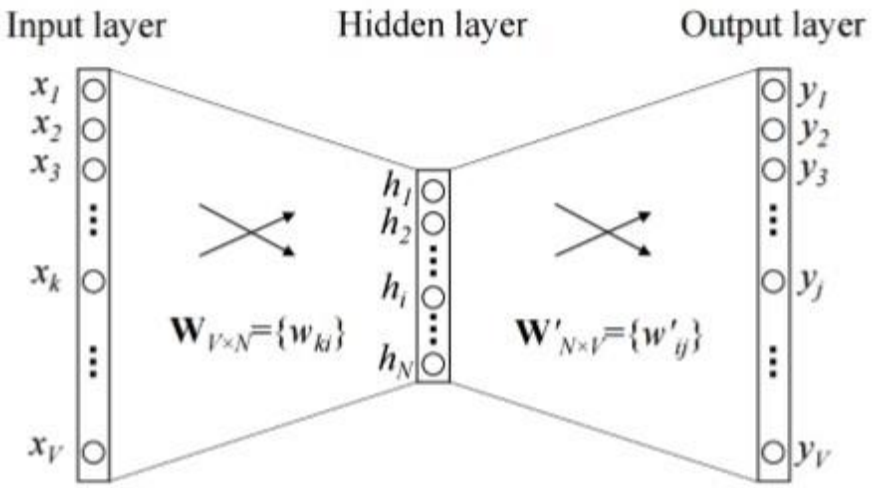

Рисунок 2.2 Модель CBOW

Word2vec имеет две разные версии: непрерывный пакет слов и Skip-Gram(непрерывный пропуск грамм). Структура изменения нейронной сети с различными версиями. На рис. 2.2 Я показываю модель CBOW, в которой контекст представлен несколькими словами для заданных целевых слов, в то время как модель SG переворачивает использование целевых слов и контекстных слов.

2.3 Вопросы конфиденциальности

Некоторые встроенные средства поиска делают сохраненные личные файлы, включая журналы электронной почты и чатов AOL, доступными для просмотра в веб-браузере, что может вызвать затруднения, если кто-то еще имеет доступ к компьютеру. А некоторые инструменты также позволяют выполнять поиск по недавно просмотренным веб-сайтам, что вызывает опасения по поводу конфиденциальности, особенно для пользователей общих компьютеров. Настольный инструмент Microsoft не индексирует и не разрешает поиск недавно просмотренных веб-сайтов, хотя и не исключает возможности делать это в будущем. YDS не индексирует кеш браузера, историю браузера или файлы избранного. Кроме того, инструмент Microsoft ищет информацию по каждому пользователю, вошедшему в систему. Если один человек использует компьютер для личных банковских операций, следующий человек, вошедший в эту машину, не сможет получить доступ к конфиденциальным данным. Технология поисковых систем в Интернете и на настольных компьютерах нуждается в радикальных изменениях, чтобы стать по-настоящему полезной. «В Интернете, когда пользователь вводит последовательность ключевых слов, даже с расширенными возможностями

поиска по ключевым словам, он может получить страницу, сообщающую ему, что существует миллион файлов, соответствующих требованиям». «Поиск на рабочем столе не намного лучше. Они дают несколько сотен или несколько тысяч. Что нужно, так это что-то более тонкое и способное более точно определить, что вы ищете».

Альтернатива, которую я предпочитаю и которая дает надежду не отставать от роста и увеличивающегося разнообразия информации как на рабочем столе, так и в Интернете, — это борьба с данными, поиск умных способов добавления метаданных и поиск лучших способов.

Заключение

На данный момент изучена предметная область и стек технологий, сделан обзор, сформулирована цель и поставлены задачи.

Список литературы

1)"What do you do for desktop search in VDI and RDSH?". Blogpost by Brian Madden on brainmadden.com.

(https://www.techtarget.com/searchvirtualdesktop/definition/virtual-desktop-infrastructure-VDI) Retrieved on March 25, 2015.

2)Anthony Ha (2 June 2008). "Lookeen offers a new way for Outlook users to search". VentureBeat. (https://venturebeat.com/2008/06/02/lookeen-offers-a-new-way-way-for-outlook-users-to-search/) Retrieved 8 March 2016.

3)Robert L. Mitchell (8 May 2013). "X1 rises again with Desktop Search 8, Virtual Edition".(https://www.computerworld.com/article/2475293/desktop-apps-x1-rises-again-with-desktop-search-8-virtual-edition.html) Retrieved 24 June 2015.

4)"Security special report: Who sees your data?", Computer Weekly, 2006-04-25. (https://www.computerweekly.com/feature/Security-special-report-Who-sees-your-data)

5)"BBC NEWS - Technology - Search wars hit desktop computers". bbc.co.uk. 26 October 2004.(http://news.bbc.co.uk/2/hi/technology/3952285.stm) Retrieved 24 June 2015.

6)"KMWorld - The Evolution of Desktop Search".

(https://www.kmworld.com/Articles/Editorial/Features/The-evolution-of-desktop-search--Good-news-for-the-knowledge-worker-9608.aspx) Retrieved 7 January 2019..

7)"dtSearch UK Blog - Desktop Wars".(https://www.dtsearch.co.uk/the-blog/blog/2014/october/23/desktop-wars.aspx) Retrieved 8 January 2019.

8)"SearchMax". goebelgroup.com. Archived from the original on 27 December 2013.

(https://web.archive.org/web/20131227130749/http://goebelgroup.com/searchtoolblog/2007/06/20/microsoft-agrees-to-change-vista-desktop-search-tool/) Retrieved 24 June 2015.

9)"Everything Search Engine". voidtools.(https://www.voidtools.com/) Retrieved 27 December 2013.

10)"Vegnos". Vegnos.(http://www.vegnos.com/) Retrieved 27 December 2013.

11)Niall Kennedy (17 October 2006). "The current state of video search". Niall Kennedy.(http://www.niallkennedy.com/blog/archives/2006/10/video-search.html) Retrieved 24 June 2015.

12)Niall Kennedy (15 October 2006). "The current state of audio search". Niall Kennedy.(https://www.niallkennedy.com/blog/2006/10/audio-search.html) Retrieved 24 June 2015.

13)"Indexing Service". microsoft.com. Microsoft.

(https://docs.microsoft.com/ru-ru/previous-versions/windows/desktop/indexsrv/indexsrv-portal?redirectedfrom=MSDN) Retrieved 24 June 2015.

14)Eduardo casais. "Converter of current to real US dollars - using the GDP deflator". areppim.com.(http://stats.areppim.com/calc/calc_usdlrxdeflator.php) Retrieved 24 June 2015.

15) Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2 [cs.CL].

16)"Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing". Google AI Blog. Retrieved 2019-11-27.(https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html)

17)Rogers, Anna; Kovaleva, Olga; Rumshisky, Anna (2020). "A Primer in BERTology: What We Know About How BERT Works". Transactions of the Association for Computational Linguistics. 8: 842–866. arXiv:2002.12327. doi:10.1162/tacl_a_00349. S2CID 211532403.(https://aclanthology.org/2020.tacl-1.54/)