УДК 004.89

Ф. В. Краснов, И. С. Смазневич

Численная оценка влияния подокументного доступа на релевантность поисковой выдачи

Рассмотрено влияние авторизации пользователей на релевантность поисковой выдачи для различных ранжирующих функций. Экспериментально показана степень искажения поисковой выдачи при подокументном доступе к коллекции. Создана модель, позволяющая количественно оценить порог для полного пересчета весов ранжирующей функции. Предложен эффективный алгоритм перерасчета весов с меньшей вычислительной сложностью, чем полный пересчет матрицы «документ – терм».

Ключевые слова: интеллектуальный поиск; ранжирующая функция; матрица весов; search engine.

Введение

Подходы к построению систем корпоративного поиска должны учитывать специфику корпоративной сети – авторизацию пользователей, ограниченное количество источников и коннекторы для доступа к определенным источникам через прикладные программные интерфейсы. В связи с этим алгоритмы поисковых систем интернета не работают в полную силу для поиска в корпоративной сети. Пустая страница с поисковой строкой является хорошо узнаваемой, но не выполняет главной для системы корпоративного поиска функции — проводника к неизвестным сайтам сети, так как в корпоративной сети новые сайты (источники) появляются крайне редко.

Развитие поисковых систем (search engine) обусловлено взрывным ростом количества интернет-ресурсов и отсутствием эффективных способов навигации по ним. В настоящее время объем данных в корпоративных сетях также растет быстро и неуклонно, поэтому поисковые системы, способные находить нужные данные в разных хранилищах информации компании, предоставляя пользователям «единое окно», становятся всё актуальнее.

На ранних этапах основной вектор развития корпоративных сетей [1, 2] предполагал копирование системообразующих принципов развития публичного сегмента интернета. Образно говоря, в корпоративной сети для сотрудников появлялись видео- и фотохостинги, корпоративные социальные сети, корпоративные мессенджеры. Следуя этому направлению, в корпоративной сети создавались поисковые системы, построенные по тем же принципам, что и в интернете, но для корпоративных ресурсов.

Но этот период в развитии корпоративных поисковых систем длился не долго. Выяснилось, что в корпоративной среде есть особенности, которые необходимо учитывать при создании поисковых систем. Одно из главных отличий корпоративной сети – это авторизация пользователей при доступе к данным. Сотрудники компании не могут анонимно посещать рабочие информационные ресурсы, так как основополагающей составляющей корпоративной системы информационной безопасности является аутентификация. Вопросам организации хранения данных для эффективной реализации моделей доступа к документам в многопользовательских системах посвящено много исследований (такие как [3, 4]), однако вопросам организации поиска в такой модели уделяется мало внимания.

Актуальность проблемы подтверждается интересом корпораций к решениям, способным обеспечить поиск по данным с различным уровнем доступа, этот запрос рынка авторы наблюдают в процессе своей аналитической работы. В существующих системах вопрос решается преимущественно разграничением поиска по разным хранилищам. В то же время очевидна необходимость в инструментах поиска по нескольким источникам корпоративных данных с единой страницы, и в этом случае возникает проблема учета прав доступа пользователей при формировании поисковой выдачи. Первый вариант — показывать в поисковой выдаче только те данные, которые доступны данному пользователю; второй — показывать в выдаче все найденные документы, но не давать пользователю возможности видеть содержимое закрытых для него файлов. Обе схемы имеют право быть рассмотренными при выборе архитектуры корпоративной информационной системы. Однако во втором случае предполагается отсутствие критически важной информации в названиях документов, что не всегда может быть приемлемо. Первый вариант, таким образом, является более универсальным, гарантирующим полное соблюдение ограничений доступа.

При ограничении доступа найденные документы по-прежнему должны быть отсортированы по степени соответствия поисковому запросу — релевантности. Она определяется весом документа, который ему присваивает ранжирующая функция. Она должна брать в расчет, что, возможно, наиболее релевантные документы, найденные при обзоре всей коллекции документов, будут недоступны конкретному пользователю из-за ограничения его прав. Вариант отображения в поисковой выдаче только тех документов, к которым у пользователя есть доступ, в порядке убывания значения ранжирующей функции, построенной для полной коллекции, имеет право на существование, но вносит непредсказуемость, исследованию которой и посвящена данная статья.

По гипотезе авторов влияние подокументного доступа на корректность поисковой выдачи имеет пороговый характер в зависимости от количества доступных документов и частотных характеристик термов запроса. После наступления порога необходимо пересчитывать веса ранжирующей функции, и в настоящей статье для этого предлагается эффективный алгоритм.

В разделе «Описание методики» рассмотрены несколько следствий подокументного доступа и их влияние на релевантность, а также возможность уменьшения сложности пересчета весов для коллекции. В разделе «Описание экспериментов» приведены условия цифрового эксперимента и численное подтверждение разработанной авторами методики на реальной коллекции документов. В «Заключении» приведены основные результаты эксперимента, подтверждающие выдвинутую автора.

Описание методики

Поясним, в чем состоит влияние авторизации на результаты поиска. Основная суть поисковой системы состоит в том, чтобы реализовать структуру данных для быстрого нахождения документов по их составляющим – словам, словоформам, фразам, фрагментам текста и изображениям. Такая структура называется «обратным индексом» (таблица 1). Построение структуры данных «обратного индекса» — ресурсоемкая задача, основная цель которой состоит в том, чтобы время доступа к искомым файлам не росло с ростом количества индексируемых файлов.

Таблица 1

«Обратный индекс»

Терм	**Документы**
знание	d_1, d_5, d_6
сила	d_2
наказание	d_5, d_7
наука	d_3

Поэтому «обратный индекс» строится заранее для длительного использования, причем строится он в течение достаточно долгого времени по сравнению со временем отклика на поисковые запросы.

Перестраивать «обратный индекс» в зависимости от полученного поискового запроса не представляется возможным. На основании «обратного индекса» в ответ на поисковый запрос выдаются идентификаторы документов, удовлетворяющие запросу, и отсортированные в соответствии с релевантностью запросу.

Поиск осуществляется по некоторой коллекции документов. В случае интернет-поиска коллекцией является набор всех собранных веб-страниц; в корпоративном поиске такой коллекцией может быть тематическая подборка документов, например, «Нормативно-правовые акты» или «Договоры на закупки за 2020 год». С введением подокументного доступа коллекция документов начинает меняться в зависимости как от роли пользователя (наделенной некоторыми правами по отношению к документам), так и от состава поискового запроса. Отметим следствия этих изменений:

Следствие 1. Изменение словаря коллекции. Словарь формируется из уникальных термов всей коллекции D. Уменьшение D приведет к уменьшению словаря. Количество термов в словаре определяет размерность векторного пространства для матрицы «документ—терм». Это означает, что изменится размерность векторного пространства для представления коллекции.

Следствие 2. Изменение весов. Исключение из коллекции D документов, недоступных для пользователя U, означает, что изменятся веса для ранжирования термов в оставшихся документах. Это приведет к деградации релевантности поисковой выдачи.

Деградация релевантности поисковой выдачи будет зависеть от того, какие веса имеют документы, к которым у пользователя нет доступа, и каков их вклад в ранжирование. Введем следующие обозначения:

- $|D|$ – число документов в коллекции,
- $|t_d|$ – количество уникальных термов в коллекции D,
- t_D – словарь коллекции,
- $|d|$ – количество уникальных термов в документе d,
- t_d – список уникальных термов в документе d,
- n_{td} – количество раз, которое терм t встречается в документе d,
- n_{tD} – число документов из коллекции D, в которых t встречается один и более раз ($n_{td} > 1$).

Поисковый запрос может состоять из нескольких термов. Однако в данной статье ограничимся простейшим случаем, когда ищется только один терм.

Рассмотрим ранжирующую функцию $F(t,d,D)$, где t – это терм, d – документ, D – коллекция документов. Для ее вычисления возьмем за основу алгоритм TF-IDF [5] (1):

$$F(t,d,D) = TFIDF(t,d,D) = \frac{n_{td}}{|d|} * log\left(\frac{|D|}{n_{tD}}\right) \quad (1)$$

Эта функция позволяет сортировать результаты поиска по убыванию вычисленного с ее помощью веса документа в зависимости от искомого терма и коллекции документов, в которой происходит поиск.

Коллекция D состоит из документов $d_1, \ldots, d_{|D|}$. Пусть для пользователя U доступна только часть коллекции, например, первые $|D_U|$ документов $d_1, \ldots d_{|D_U|}$. Тогда вес терма t в документе d изменится согласно следствиям 1 и 2: $TFIDF(t, d, D_U)$ будет иметь нулевое значение для всех документов, недоступных пользователю, а для документов, на просмотр которых права есть, изменится значение выражения под логарифмом.

Для дальнейшего понимания изменений функции $TFIDF(t, d, D)$ в связи с подокументным доступом рассмотрим векторное представление коллекции — матрицу «документ-терм». Размерность матрицы будет составлять $|t_D| \times |D|$, где $|t_D|$ – количество уникальных термов в коллекции, а $|D|$ – количество документов в ней. Для пользователя U эта матрица будет обладать меньшей размерностью $|t_{D_U}| \times |D_U|$, а значения ее ячеек изменятся по-разному в зависимости от терма. Сформулируем еще два следствия изменения числа доступных пользователю документов.

Следствие 3. Неравномерное изменение весов: для любого терма, не входящего в недоступные пользователю документы, веса не изменятся, поскольку не поменяется число содержащих его документов.

Практическое значение следствия 3 состоит в том, что актуальная сложность вычислений O_{D_U} для $TFIDF(t, d, D_U)$ оказывается меньше, чем при полном пересчете всех значений весов.

Следствие 4. При подокументном доступе число вхождений термов в документы n_{td} не меняется в зависимости от прав доступа пользователя.

Исходя из следствий 3 и 4, предложим оптимизированный алгоритм для пересчета матрицы $TFIDF(t, d, D_U)$.

Согласно следствию 4, при введении подокументного доступа нет необходимости пересчитывать компонент ранжирующей функции n_{td}.

Если из коллекции D исключаются документы $\acute{D}_U$, недоступные пользователю, то новое значение общего числа вхождений терма, присутствующего в исключенных документах, по коллекции n_{tD_U} может быть посчитано как разность $n_{tD} - n_{t\acute{D}_U}$:

$$TFIDF(t,d,D_U) = \begin{cases} \frac{n_{td}}{|d|} * log\left(\frac{|D_U|}{n_{tD}}\right), & \forall t_{\acute{D}_U} \notin t_{D_U} \\ \frac{n_{td}}{|d|} * log\left(\frac{|D_U|}{n_{tD} - n_{t\acute{D}_U}}\right), & \forall t_{\acute{D}_U} \in t_{D_U} \end{cases} \quad (2)$$

Рассмотрим подробнее составляющие уравнения (2). Первая часть отражает практический результат следствия 3, поэтому пересчет весов сводится к уменьшению числа документов $|D_U|$. Во втором варианте, когда термы из исключенных документов встречаются в оставшихся, знаменатель

выражения под логарифмом изменится. Но это изменение, по сути, означает вычисление разницы двух векторов, а не полный пересчет всей матрицы. Таким образом, появляется методическая возможность для реализации более быстрого алгоритма.

Для тестирования преимуществ этого подхода была проведена серия экспериментов.

Описание экспериментов

Эксперименты проводились на наборе данных – коллекции из $|D| = 10^3$ текстовых документов общей тематики и разной длины, случайным образом выбранных из большей коллекции. Словарь коллекции составил порядка $|t_D| = 10^4$ термов.

Для первого и второго экспериментов в документы коллекции *D* добавлялся суррогатный терм t^s в разном количестве $\mu*|D|$, $\mu \in (0.1, \ldots, 0.9)$, так чтобы тексты содержали этот терм в различных пропорциях – от 1 до 100 штук на один текст. Таким образом, при поиске по t^s должно отображаться $\mu*|D|$ текстов, отсортированных по убыванию ранжирующей функции *F (t, d, D)*.

В первом эксперименте была проверена гипотеза о ненарушении порядка ранжирования при переходе к подокументному доступу. Из коллекции *D* случайным образом удалялось различное количество документов, как среди содержащих суррогатный терм, так и среди документов без него. На рис. 1 показана зависимость, отражающая порядок выдачи по убыванию значения ранжирующей функции.

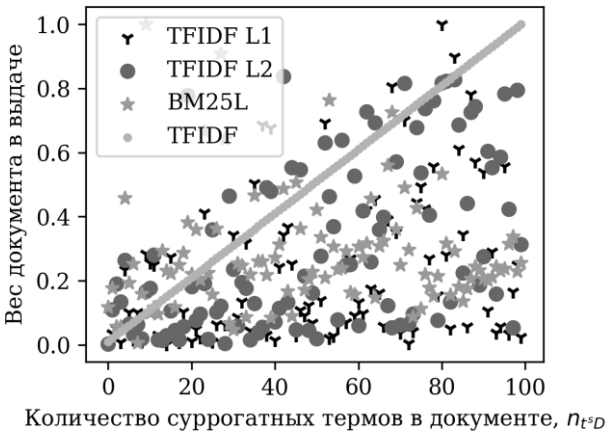

Рис. 1. Веса документов при поиске по суррогатному терму для различных ранжирующих функций

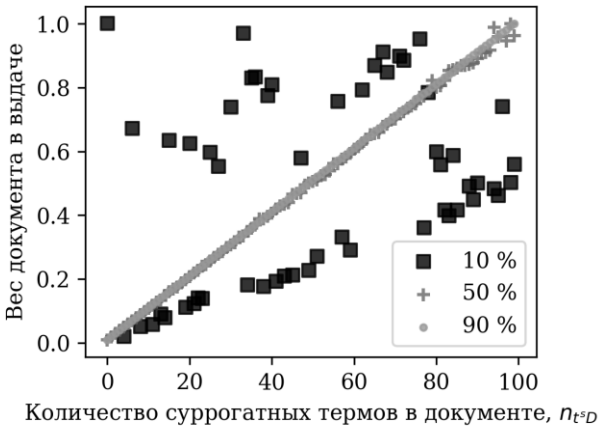

Рис. 2. Зависимость весов от количества суррогатных документов в коллекции при разном количестве доступных документов (η)

Следует отметить влияние нормализации [6] (L2, L1) на порядок выдачи. Так как документы имеют различную длину, нормирование ранжирующей функции в документе приводит к искажению порядка выдачи. Очевидно, что вместо линейной зависимости веса от количества суррогатных термов возникает разброс значений. Наиболее адекватную релевантность показывает TFIDF без нормирования, так как больший вес получают документы, содержащие большее количество суррогатных термов. В дальнейшем примем значения весов TFIDF без нормировки за базовые и будем измерять отклонения от них.

Во втором эксперименте автор создал модель подокументного доступа путем исключения из коллекции в случайном порядке определенного количества документов. Это означает, что у пользователя есть доступ только к некоторой доле η от общего числа документов коллекции, например, к 10%, 50% или 90%. По этим документам произведем полный перерасчет весов ранжирующей функции *TFIDF (t, d, D_U)*.

На рис. 2 изображены зависимости весов документов от их количества в коллекции для документов с различным числом суррогатных термов. Можно наблюдать, что при доступности η=90% коллекции веса почти не отличаются от базовых. Но с уменьшением количества доступных документов до 10% от всей коллекции порядок весов нарушается, а значит нарушается общая релевантность поисковой выдачи. Мерой нарушения релевантности поисковой выдачи будем считать метрику

Weighted Mean Absolute Error (WMAE) [7], равную сумме взвешенных отклонений весов документов от базовых значений весов. Значение метрики WMAE для конкретной ранжирующей функции будет показывать, насколько сильно изменился порядок выдачи при изменении прав доступа. Таким образом, WMAE будет обратна релевантности выдачи поисковому запросу: чем больше значение WMAE, тем менее релевантна выдача по сравнению с базовой. Метрика WMAE, в свою очередь, зависит от количества искомых термов в коллекции.

На рис. 3 изображена зависимость WMAE при разном количестве суррогатных термов в коллекции от доли доступных документов η.

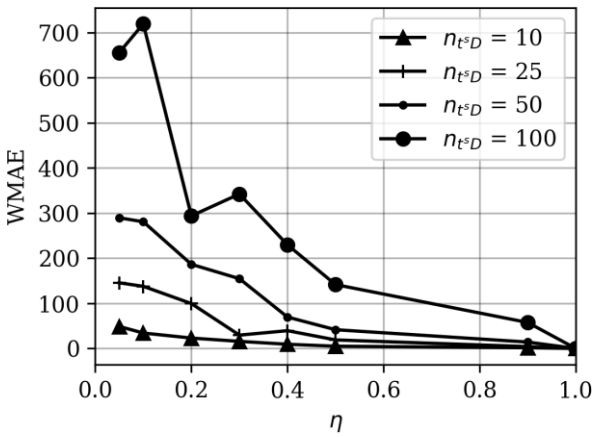

Рис. 3. Зависимость WMAE для долей доступных документов коллекции при разных количествах суррогатных термов ($n_t^s D$)

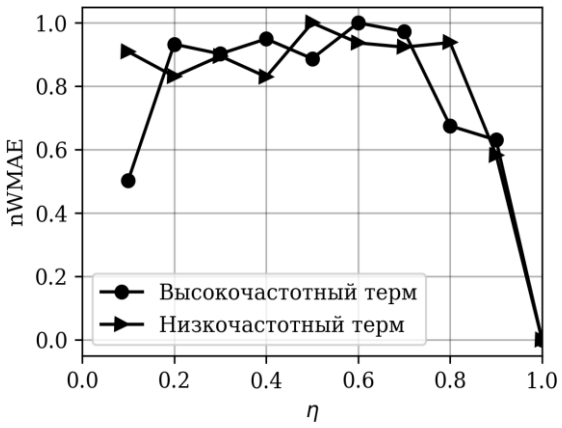

Рис. 4. Зависимость nWMAE для долей доступных документов коллекции

Подтверждение гипотезы о возникающей необходимости полного перерасчета весов основано на том факте, что значения метрики WMAE, характеризующей меру нарушения релевантности поисковой выдачи при добавлении суррогатных термов, значительно растет, когда доля доступных документов η становится менее 40 %. Также отметим, что рост WMAE происходит для термов, чаще встречающихся в коллекции.

Третий эксперимент был проведен с термами, выбранными случайным образом из коллекции *D* (без добавления суррогатных). Распределение Ципфа [8] («ранг—частота») для коллекции *D* было поделено на два промежутка: высокочастотные и низкочастотные термы (граница проходит по значению ранга терма, равному 5000, при общем словаре коллекции около 10 000). Из каждого промежутка было выбрано одинаковое количество случайных термов, произведено вычисление нормированной версии nWMAE для различных значений η, а после этого результат был усреднен. Полученная зависимость представлена на рис. 4 и показывает, что нарушения порядка документов в поисковой выдаче начинаются уже при 80% доступных документов. Это подтверждает высказанную ранее гипотезу.

Заключение

В статье изучено влияние подокументного доступа на релевантность поисковой выдачи. В результате проведенного исследования сделаны следующие выводы:

1. Порядок документов в поисковой выдаче изменяется при введении подокументного доступа (авторизации).

2. Подход с учетом применения прав доступа к отсортированной поисковой выдаче будет вводить в заблуждение пользователя, так как наиболее релевантные поисковому запросу документы не будут отображаться первыми.

3. Для повышения релевантности поисковой выдачи при использовании подокументного доступа необходимо производить перерасчет весов ранжирующей функции.

Эффект нарушения порядка выдачи зависит от доли документов, доступных для пользователя, и близок к обратному закону: чем меньше документов из коллекции доступно, тем больше нарушений в

релевантности поисковой выдачи. Также эффект нарушения выше по абсолютному значению метрики WMAE для высокочастотных термов. Для редких термов эффект нарушения релевантности выдачи ниже, а значит уменьшается и необходимость полного пересчета. Полный пересчет весов коллекции необходим при снижении доли доступных документов до 80% и при поисковых запросах с термами, число документов с которыми превышает 10% от общего числа документов коллекции.

Предложенный алгоритм показывает теоретические возможности по ускорению расчетов по сравнению с полным пересчетом всех значений весов. На практике алгоритм реализуется в рамках проекта по разработке платформы интеллектуального управления и семантического анализа данных Naumen LegalTech, нацеленной на применение перспективных научно-обоснованных решений в реальных информационных системах для российских корпоративных пользователей.

В заключение стоит отметить, что рассмотренные в статье алгоритмы имеют применение и в смежных областях, например, в рекомендательных системах, когда неточный порядок рекомендаций может привести к критичным последствиям.

ЛИТЕРАТУРА

1. Шелупанов А. А., Елапин К. А., Максименко А. Стратегическое планирование защищенных корпоративных компьютерных сетей //Интеллектуальные системы в управлении, конструировании и образовании. 2001. С. 62-71.
2. Богданов В. Н., Вихлянцев П. С., Симонов М. В. Построение корпоративной сети на базе сети АТМ общего пользования //ИНФОРМОСТ"–" Радиоэлектроника и Телекоммуникации. 2003. №. 5. С. 29.
3. Лизин С. Управление данными в корпоративных системах //Открытые системы. СУБД. 2010. №. 8. С. 31-31.
4. Галлямшин В., Скок А. Когда защита не работает //Открытые системы. СУБД. 2014. №. 9. С. 36-37.
5. Jones K. S. A statistical interpretation of term specificity and its application in retrieval // Journal of documentation. 1972. Т. 28. №. 1. С. 11-21.
6. Singhal A., Buckley C., Mitra M. Pivoted document length normalization //Acm sigir forum. New York, NY, USA : ACM, 2017. Т. 51. №. 2. С. 176-184.
7. Cleger-Tamayo S., Fernández-Luna J. M., Huete J. F. On the Use of Weighted Mean Absolute Error in Recommender Systems //RUE@ RecSys. 2012. С. 24-26.
8. Zipf G.K. Human Behavior and the Principle of Least Effort. Addison-Wesley Press, 1949. С. 484-490. 573 с.