DOI: 10.17323/2587-814X.2021.2.60.74

Проблема потери решений в задаче поиска схожих документов: Применение терминологии при построении векторной модели корпуса

Ф.В. Краснов
E-mail: fkrasnov@naumen.ru

И.С. Смазневич
E-mail: ismaznevich@naumen.ru

Е.Н. Баскакова
E-mail: enbaskakova@naumen.ru

NAUMEN R&D
Адрес: 620028, г. Екатеринбург, ул. Татищева, д. 49А

Аннотация

В статье рассматривается задача поиска схожих по смыслу текстовых документов в корпусе. Исследуется проблема невыявления алгоритмом TF-IDF части решений, возникающая при разработке прикладных интеллектуальных информационных систем: потеря пар, схожих согласно человеческой оценке, но получающих низкую оценку схожести от программы. Предложена модификация алгоритма с заменой общего словаря на словарь специализированных терминов. Добавление тезаурусов при построении векторной модели корпуса, основанной на ранжирующей функции, не было ранее исследовано; применение тезаурусов до сих пор изучалось лишь для улучшения тематической модели. Цель работы – повысить качество решения, минимизируя потерю значимой его части и не добавляя «ложно-схожие» пары документов, за счет применения при векторном разложении TF-IDF словаря терминов, выделенного из текста анализируемых документов. Эксперимент проведен поочередно на двух корпусах структурированных нормативно-технических документов, объединенных тематически: стандартов в отношении информационных технологий и в сфере железных дорог. Словарь терминов составлен при автоматическом анализе текста рассматриваемых документов методами выделения именованных сущностей, основанных на правилах. Продемонстрировано, что разложение TF-IDF по словарю терминов дает больше релевантных результатов для исследуемой задачи, что подтвердило выдвинутую гипотезу. Предложенный метод в меньшей степени зависит от недостатков текстового слоя (таких как ошибки распознавания), чем расчет близости документов по полному словарю корпуса. Определены факторы, способные повлиять на качество решения: способ составления словаря терминов, выбор диапазона *n*-грамм для словаря, корректность формулировки терминов и обоснованность их включения в глоссарий документа. Полученные выводы могут использоваться при решении прикладных задач, связанных с поиском близких по смыслу документов, таких как семантический поиск с учетом предметной области, корпоративный поиск в многопользовательском режиме, обнаружение скрытого плагиата, выявление противоречий в коллекции документов, определение новизны в документах при построении базы знаний.

Ключевые слова: схожесть документов; семантическая близость документов; применение тезаурусов; векторная модель корпуса; прикладные интеллектуальные информационные системы; объяснимость алгоритма; оценка схожести; интеллектуальный анализ текста.

Цитирование: Краснов Ф.В., Смазневич И.С., Баскакова Е.Н. Проблема потери решений в задаче поиска схожих документов: Применение терминологии при построении векторной модели корпуса // Бизнес-информатика. 2021. Т. 15. № 2. С. 60–74. DOI: 10.17323/2587-814X.2021.2.60.74

Введение

Среди задач, которые отличают интеллектуальные прикладные информационные системы от систем автоматизации бизнес-процессов, имеет место задача обнаружения инсайтов в большом объеме информации, в частности, в базе текстовых документов компании. Эта задача включает в себя поиск «близких по смыслу» документов. Для решения данной задачи строится семантическая модель текстового корпуса, в рамках которой схожесть документов определяется как расстояние между векторами документов.

Одна из проблем заключается в возможной потере части пар документов, которые схожи с точки зрения человеческой оценки, но не удовлетворяют условию превышения порогового значения схожести, установленного в прикладной интеллектуальной информационной системе. Это приводит к задаче обнаружения той части результатов, которые не показывают достаточной степени схожести при существующих методах, но должны быть учтены с точки зрения эксперта – пользователя системы («истинно-схожих» пар документов).

При расчете векторной модели корпуса текстовых документов используются разные словари, от характеристик и ограничений которых зависит качество решения. В частности, влияние оказывают степень соответствия (сфокусированность) словаря предметной области корпуса, доля часто употребляемых и редких слов, выбор диапазона *n*-грамм и другие параметры.

Следует отметить, что при слишком сильном расширении словаря или чрезмерном снижении порогового значения схожести с целью включить в множество решений указанные «истинно-схожие» пары документов результат снова ухудшается. Это объясняется тем, что вместе с возвратом потерянной части решений в их число включаются и лишние, «ложно-схожие» решения, которые обнаруживают близость за счет малозначимой части словаря (слов с низким весом для семантики в рамках данного корпуса). Одним из методов, позволяющих добиться баланса между включением в решение лишних пар документов и потерей существенной части результатов, является использование тезаурусов предметной области при построении модели текстового корпуса.

Многие прикладные задачи требуют вычисления показателей сходства между экземплярами текстов и их составных частей – параграфов, предложений. Наиболее очевидный пример – когда пользователь ищет информацию в системе и поисковая система сопоставляет текст запроса с текстами ранее сохраненных документов, чтобы выдать наиболее релевантный документ. Запрос пользователя представляет собой короткий текст, и система выводит наиболее схожие по тексту документы с помощью функции ранжирования.

Кроме модулей семантического поиска, содержательная схожесть документов используется в следующих ИТ-решениях:

- ✦ рекомендательная система для авторов, определяющая наиболее подходящий журнал для публикации;
- ✦ система маршрутизации входящих заявок, подбирающая эксперта в соответствии с ранее обработанными им документами;
- ✦ программный модуль для формирования проектных команд по определенному техническому заданию;
- ✦ компонент СЭД для определения пути согласования документа на основании его содержания.

Для построения матрицы схожести необходимо использовать векторное представление документов, которое может быть построено с помощью статистической меры TF-IDF, часто применяемой и в качестве функции ранжирования. Для вычисления меры TF-IDF учитывается весь словарь корпуса, поэтому доминирующее влияние на близость документов может оказывать общеупотребительная лексика, в то время как отраслевая специфика документов может оказаться утерянной. Возникает дополнительная задача по выявлению документов, близких за счет специфической терминологии. Модификация

метода может быть произведена за счет изменения набора термов, на основании которых определяется сходство документов. В частности, количество и точность полученных результатов могут быть улучшены путем встраивания тезауруса предметной области в алгоритм вычисления функции ранжирования.

Такая модификация алгоритма улучшает качество рекомендаций в информационных системах и ускоряет принятие решений пользователем. Это осуществляется за счет повышения объяснимости алгоритма, то есть сокращения времени, которое необходимо пользователю для понимания, почему система рекомендует ему некоторую пару документов как похожую [1]. В конечном счете это улучшение способствует росту доверия пользователя к рекомендациям системы и упрощает его аналитическую работу с текстовыми документами – поиск информации в корпоративных источниках, проверку на отсутствие дублей в базе, обнаружение пересечений и противоречий. Все это в совокупности приводит к сокращению временных затрат сотрудников на обработку больших объемов неструктурированных данных.

Целью настоящей работы является улучшение модели на основе ранжирующих функций за счет использования в качестве тезауруса словаря терминов предметной области. Авторы сфокусировались на корпусах структурированных текстовых документов в определенной предметной области. К таким документам, например, относятся нормативно-технические базы организаций [2], доходные и расходные договоры, резюме кандидатов, ГОСТы и многие другие. Рассматривалась прикладная задача нахождения в текстовом корпусе схожих по смыслу документов.

Исследовалась проблема потери части решения при расчете близости документов с помощью ранжирующей функции. Данная проблема состоит в том, что некоторые пары документов являются схожими согласно человеческой оценке, однако программа не определяет их как таковые, поскольку они показывают низкую степень схожести по алгоритму TF-IDF, даже с учетом оптимизации словаря за счет удаления наиболее часто и наиболее редко встречающихся слов. На практике такая проблема возникает при разработке прикладных интеллектуальных информационных систем, где ставится задача найти в корпусе документов достаточно близкие, то есть такие, степень схожести которых превышает некоторый установленный порог и, следовательно, является значимой для пользователя. В работе предложена модификация существующих методов вычисления близости документов с помощью ранжирующих функций, позволяющая избежать потери указанной части результатов.

Гипотеза исследования состоит в том, что при поиске в текстовом корпусе пар близких по смыслу документов с помощью ранжирующей функции TF-IDF та часть решения, которая теряется при векторном разложении по полному словарю корпуса, может быть обнаружена при построении векторной модели на основании словаря терминов, соответствующих предметной области. На *рисунке 1* показано графическое представление гипотезы.

Статья включает обзор имеющихся исследований в области применимости тезаурусов и проблематики их построения, описание методов исследования и экспериментального подтверждения исследовательской гипотезы, а также анализ полученных результатов.

1. Применение тезаурусов

Компьютерная лингвистика десятилетиями интересовалась возможностью применения тезаурусов при анализе текстов для выявления семантических связей в корпусе, поскольку созданные экспертами словари определений терминов и понятий обладают высокой смысловой точностью.

Рис. 1. Ожидаемый сдвиг графика распределения степени схожести документов при замене в алгоритме TF–IDF общего словаря на словарь терминов предметной области

Первые методы, предложенные в 70-е и 80-е годы XX века, были нацелены на создание семантических языковых моделей с помощью анализа компьютерных словарей – источников таксономических отношений «гипоним – гипероним». Для извлечения отношений из словарей использовались алгоритмы, основанные на правилах [3, 4]. Преимущество этих методов состояло в том, что семантические отношения извлекались ими из надежных источников, которые можно считать уже частично структурированными, поскольку словари работают как «имплицитные таксономии». Однако эти методы унаследовали проблемы лексикографического материала, связанные как с возможной недостоверностью данных вследствие некорректного обновления источников, так и с неполнотой словарей, поскольку даже сегодня они не всегда являются корпусными.

Во многих исследованиях проявлялся интерес к поиску со-гипонимических отношений, то есть выявлению групп слов, которые определяются с помощью одного и того же гиперонима, таких как типы систем, оборудования, связей [5, 6]. Такие слова считаются парадигматически связанными. Это означает, что они имеют тенденцию повторяться в сходных синтагматических контекстах и предположительно обладают общими семантическими признаками.

Другая стратегия выявления семантических отношений в корпусе состоит в создании языковой модели на основе статистического анализа текстов без использования предварительно полученных знаний о связях словаря. В работе [7] для выделения пар гипоним-гипероним строится направленный граф совстречаемости термов, а если они слишком редкие, то в частотное распределение внедряется смещение: вводится отношение аналогии между гипонимами на основе состава и морфологических признаков термов. Например, если статистически установлено, что синдром Аспергера и синдром Карпентера (гипонимы) – заболевания (гипероним), то терм «синдром Меретойи» имеет тот же гипероним «заболевание»; к нему же может быть отнесено и слово «артрит» по аналогии с «гастритом». Авторы статьи [8] соединяют обе стратегии для создания гипернимических цепочек с использованием подходов, основанных на дистрибутивной семантике.

Потенциальная ограниченность количественных методов состоит в недостатке точности, однако проблема решается путем увеличения объема лингвистических данных. Поэтому такой подход считается эффективным и является довольно популярным. Кроме того, будучи независимым от языка, он может быть легко реплицирован и использован для создания многоязычных ресурсов.

В рамках статистического подхода для решения многих прикладных задач строятся векторные модели корпуса документов на основании ранжирующих функций. Часто используемая функция ранжирования TF-IDF, которая основана на двух понятиях: частота появления слова (терма) в документе (term frequency, TF) и важность этого слова для всего набора документов – обратная частота вхождения слова во все документы корпуса (inverse document frequency, IDF). Метод был предложен в 1970-х годах [9–12] и до сих пор широко применяется для анализа схожести текстов, поскольку позволяет сохранить определенные их свойства при проекции текстовых данных на числовое пространство. Измерение схожести может быть сделано через вычисление близости между TF-IDF векторами, например, по косинусной мере. Принцип сравнения с помощью TF-IDF можно назвать «пословным», так как в разряженном TF-IDF векторе число компонент соответствует числу термов в словаре. Однако набор термов, определяющих сходство пары документов, является результатом частотной дискриминации, а не содержания этих документов. Ранжирование на основе TF-IDF используется в различных приложениях, таких как кластеризация документов [13–16] и тематическое моделирование [17, 18].

Использование тезаурусов при анализе корпуса текстов методами дистрибутивной семантики рассмотрено в ряде исследований. В частности, в разное время авторы обращались к проблематике работы с тезаурусом на одном из шагов алгоритма в задаче текстового поиска. Например, в статье [19] описан подход к автоматической индексации тезисов статей с использованием тезауруса предметной области. Авторами работы [20] представлен способ автоматического извлечения ключевых фраз с помощью семантической информации о терминах и фразах, собранных из предметно-ориентированного тезауруса. В работе [21] применяется политематический тезаурус для целей семантического сопоставления понятийных образов.

Заметное число исследований посвящено улучшению параметров тематических векторных моделей корпуса, в которых векторное разложение документов сделано по набору тем, выделенных в

результате анализа текста на всем корпусе. Например, в работе [22] улучшение тематической модели производится с помощью искусственного увеличения совстречаемости синонимов, а для усиления когерентности тем модели в исследовании [23] информация о синонимах вносится в априорное распределение Дирихле. В работе [24] предложено такое понятие как Thesaurus-Based Topic Model и произведено сравнение различных тематических моделей. Все вышеперечисленные исследования объединяет общий результат экспериментов: исключение гипонимов и увеличение частот фраз улучшает человеческую оценку качества получаемых тем, поскольку придает больший вес более конкретным словам и фразам, содержание которых лучше определено.

В то же время возможность применения тезаурусов для улучшения семантической модели корпуса текстов, в которой матрица близости строится на основании ранжирующей функции (такой как TF-IDF), до сих пор не была достаточно изучена.

2. Построение тезаурусов

Согласно определению Международной организации по стандартизации (ISO)[1], тезаурус является словарем, формально организованным для того, чтобы установить явные априорные отношения между понятиями. Элементами тезауруса являются лексические единицы и семантические отношения (связи) между ними. Тезаурусные отношения (род – вид, часть – целое, комплекс – элемент, причина – следствие) налагаются на структуру таксономии, то есть идентифицируются как основные таксономии предметной области.

Методика написания тезаурусов определена в стандарте «ГОСТ Р ИСО 704-2010 Терминологическая работа. Принципы и методы» [25]. Исторически тезаурусы создавались для ручного индексирования документов и при их создании не принимались во внимание вопросы, связанные с автоматической индексацией. Трудность построения тезауруса, соответствующего всему тематическому многообразию индексируемой информации, делает тезаурус самодостаточным информационным продуктом, одновременно являясь основной причиной его непопулярности в современных информационных системах.

В качестве выдающегося примера специализированных тезаурусов следует отметить российский тезаурус в области сельского хозяйства, созданный Центральной научной сельскохозяйственной библиотекой [26]. Тезаурус построен как расширение словаря-классификатора ГРНТИ (Государственного рубрикатора научно-технической информации).

Трудоемкость ручного составления тезаурусов и возникающие вследствие этого проблемы можно увидеть на примере таких нормативно-технических документов, как ГОСТы. Текст большинства из них сопровождается тезаурусом – разделом с описанием терминов и определений. Организации, отвечающие за разработку отраслевых стандартов, сталкиваются с тем, что зачастую в разных документах обнаруживаются противоречащие определения одних и тех же терминов, разные названия одного понятия или конфликты между нормативными значениями показателей. Причина этого явления состоит в том, что база ГОСТов, состоящая из сотен тысяч документов, накапливалась в течение многих десятилетий. При этом из-за ограниченных возможностей поиска положения стандартов зачастую дублировались, превращаясь со временем в противоречия из-за обновления документов без учета их тематических связей с другими.

Пример 1. Рассмотрим различные определения и названия понятия «акустический (шумозащитный) экран», имеющиеся в разных нормативно-технических документах:

1. ГОСТ Р 51943-2002. Экраны акустические для защиты от шума транспорта. Методы экспериментальной оценки эффективности (2002): «**Акустический экран**, **экран**: Барьер (ограниченная преграда), устанавливаемый на пути распространения шума реального источника к защищаемому от шума объекту».

2. ГОСТ 32957-2014. Дороги автомобильные общего пользования. Экраны акустические. Технические требования (2014): «**Акустический экран**: Искусственная преграда, устанавливаемая на пути распространения шума от автомобильного транспорта к защищаемому от шума объекту. Типовой акустический экран представляет собой сборную конструкцию, состоящую из следующих основных частей: фундамента (если предусмотрено проектной документацией), несущей конструкции (в частности, опорных стоек) и панелей. В качестве

[1] http://docs.cntd.ru/document/1200129056 .

дополнительных элементов используют уплотнения, поперечные профилированные балки, крепежные детали, акустические развязки, козырьки, калитки, ворота, рамы разрывов и т.п.».

3. ГОСТ 33329-2015. Экраны акустические для железнодорожного транспорта. Технические требования (2015): «**Акустический экран**: Протяженная искусственная преграда, устанавливаемая на пути распространения шума от реального источника (железнодорожного транспорта) к защищаемому от шума объекту».

4. СП 338.1325800.2018. Защита от шума для высокоскоростных железнодорожных линий. Правила проектирования и строительства (2018): «**Шумозащитный (акустический) экран (экран); ШЭ**: Сооружения в виде вертикальных или наклонных протяженных искусственных преград различной конструкции, земляных насыпей, выемок, галерей и т.п., устанавливаемых вдоль железных дорог на пути распространения шума транспортного потока к защищаемому объекту в целях снижения шума».

5. ОДМ 218.8.011-2018. Методические рекомендации по определению характеристик и выбору шумозащитных конструкций автомобильных дорог (2018): «**Шумозащитный экран (ШЭ)**: Протяженная искусственная преграда, устанавливаемая на пути распространения шума от автомобильного транспорта к защищаемому от шума объекту, ширина (или толщина) которой много меньше ее высоты, состоящая из фундамента и закрепленного на нем шумозащитного полотна».

Чтобы избежать влияния указанных недостатков тезаурусов, составленных вручную, на результаты поиска схожих документов в прикладных информационных системах, необходимо использовать автоматические методы формирования терминологических словарей для корпуса документов.

Существующие подходы к извлечению семантических отношений на основе паттернов, подобные описанным в предыдущем параграфе, используются и в вычислительной терминологии. Например, в работе [27] описаны методы экстракции, ориентированные на выявление отношений гипоним – гипероним (частное и обобщение), мероним – холоним (часть и целое), синонимии и причинно-следственных связей, которые совокупно используются для определения терминов и их отношений в рамках подхода “data fusion pipeline”. Однако вычислительная терминология сосредоточена в основном на изучении самих паттернов семантических отношений, их описании, интерпретации и формализации их лингвистических свойств, а также на анализе паттернов за пределами возможностей их обнаружения. Однако для рассматриваемой задачи достаточным является получение набора терминов, которые используются в исследуемом корпусе документов без учета семантических взаимоотношений этих терминов.

3. Методы

Для проверки сформулированной гипотезы был проведен эксперимент с целью сравнения результатов решения задачи поиска схожих документов в корпусе, полученных двумя методами – базовым TF-IDF и его модификацией с применением тезауруса.

В рамках исследования был рассмотрен набор из N_D структурированных текстовых документов, объединенных тематическим направлением.

В качестве тезауруса (как наиболее доступный вариант) был использован словарь T из терминов предметной области, полученный в результате автоматического анализа текстов этих документов. Для этого все документы из набора были обработаны алгоритмом извлечения именованных сущностей, основанным на правилах (rule-based NER). При этом из соответствующего раздела каждого документа, где перечислены используемые термины и их описания, были извлечены списки терминов, которые затем были объединены в словарь терминов, специфичных для данного корпуса.

В словарь включались только термины, состоящие из одного или двух слов. Слова были приведены к нормальной форме, внесена информация о морфологических признаках, исключены цифры и английские символы.

Документы были преобразованы к следующему текстовому формату: набор документов представлен в виде массива строк M, в котором одна строка соответствует одному документу. Последовательность слов при этом была сохранена.

Для получения матриц «пословной» степени схожести S_{TFIDF} текстовых документов корпуса M было использовано векторное преобразование TF-IDF в нескольких вариантах, различающихся словарем:

✦ в качестве словаря были использованы все слова из корпуса документов. Было рассмотрено не-

сколько вариантов построения словаря в зависимости от пороговой частоты использования слов в рамках всего корпуса: если слово встречалось менее определенного количества раз (параметр df_{min}), то оно исключалось из словаря;

✦ в качестве словаря был использован набор специализированных терминов *T*.

В результате для каждого массива было получено пять вариантов матрицы $\mathbf{A}_{TFIDF}$ типа «документы – словарь» с размерностью $N_D \times N_T$ (число слов в словаре). Матрица близости векторов (степени схожести документов) $\mathbf{S}_{TFIDF}$ вычислялась по формуле:

$$\mathbf{S}_{TFIDF} = \mathbf{A}_{TFIDF} \times \mathbf{A}^{T}_{TFIDF}.$$

На основании матрицы $\mathbf{A}_{TFIDF}$, сформированной по набору специализированных терминов, был построен граф знаний о корпусе текстовых документов $\mathbf{G}_{TFIDF}$. Вершины графа (документы и термины) соединялись ребрами в том случае, если значение TF-IDF разложения для соответствующей пары «документ – термин» оказывалось выше заданного порога (вес ребра равен соответствующему значению в матрице $\mathbf{A}_{TFIDF}$).

4. Результаты эксперимента

Для проведения эксперимента был выбран набор юридически значимых документов из библиотеки ГОСТ: $N_{D\_IT} = 667$ документов в формате docx, относящихся к тематике информационных технологий. После обработки документов в соответствии с выбранной методикой был получен массив текстовых строк M_{IT}. Из текста документов было выделено $T_{IT} = 4417$ терминов.

Результаты вычислений для документов по информационным технологиям показаны на *рисунке 2.*

В *таблице 1* приведены значения математического ожидания и среднеквадратического отклонения степени схожести пар документов M_{IT}.

Таблица 1.

Математическое ожидание и среднеквадратическое отклонение степени схожести пар документов M_{IT} при различных словарях

Условия	Математическое ожидание, %	Среднеквадратическое отклонение, %
По полному словарю при $df_{min} = 0$	2,9	2,3
По полному словарю при $df_{min} = 5$	5,8	3,9
По полному словарю при $df_{min} = 10$	7,9	4,8
По полному словарю при $df_{min} = 50$	17,4	8,0
По словарю терминов	21,9	8,0

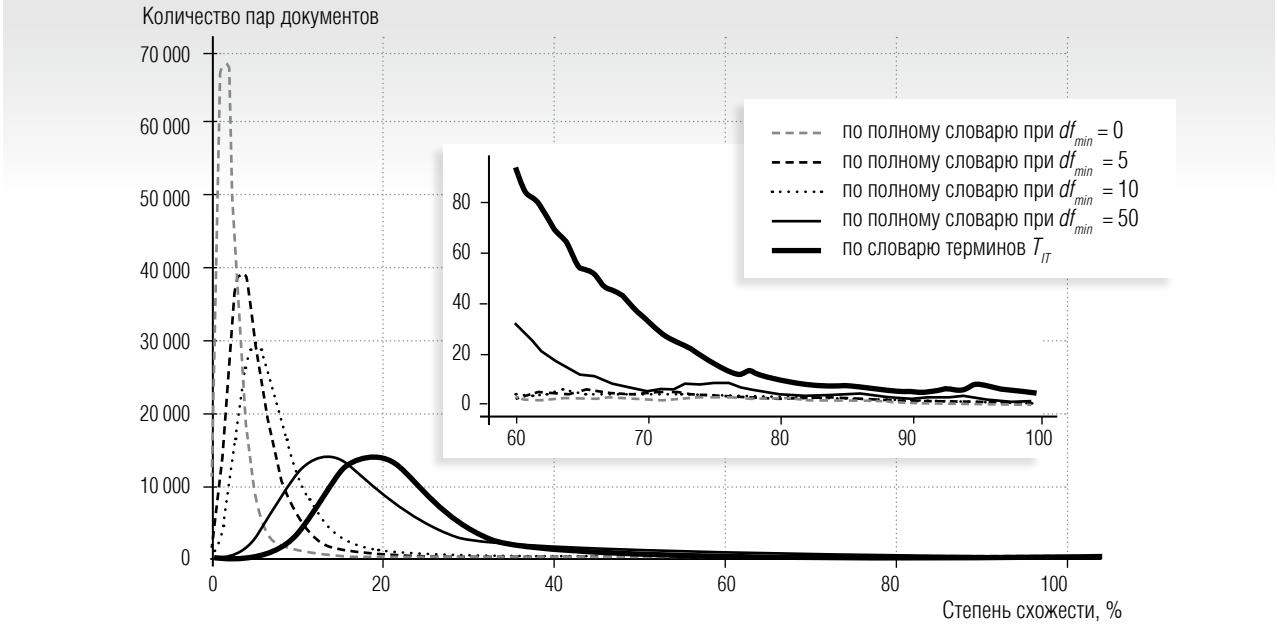

Рис. 2. Количество пар документов M_{IT} разной степени схожести при различных вариантах словаря, в т.ч. количество пар документов M_{IT} со схожестью в диапазоне 60 – 100% при разных вариантах словаря

Эксперимент подтвердил выдвинутую гипотезу: замена общего словаря на набор специализированных терминов в алгоритме на основе ранжирующей функции увеличивает количество найденных решений. На представленных графиках и из данных таблицы видно, что построение модели на специализированных терминах определяет большее число пар схожих документов, а варианты с исключением редко встречающихся в рамках корпуса слов существенных отличий не вносят. Поэтому для дальнейшего сравнения методов, включая анализ распределения близости документов, из первой группы вариантов алгоритма был выбран один – с расчетом схожести по полному словарю при $df_{min} = 5$.

Для предметного анализа был рассмотрен диапазон схожести от 60% до 100%. Количественные характеристики полученного результата представлены в *таблице 2.*

Пример 2. В *таблице 3* приведены примеры пар документов по информационным технологиям, имеющих высокую степень близости TF-IDF по словарю терминов и низкую – по полному словарю корпуса.

Таблица 2.

Количество пар документов M_{IT} в разных интервалах степени схожести, обнаруженное алгоритмом TF-IDF с различными словарями

Степень близости документов M_{IT}	Количество пар близких документов, определенное по полному словарю при df_{min}= 5	Количество пар близких документов, определенное по словарю терминов
60% – 70%	42	459
70% – 80%	26	133
80% – 90%	25	55
90% – 100%	7	56

Таблица 3.

Примеры пар документов M_{IT}, имеющих высокую степень близости TF-IDF по словарю терминов и низкую – по полному словарю корпуса

№ пары	Документы	Абсолютное значение близости по TF-IDF по словарю терминов	Абсолютное значение близости по TF-IDF по полному словарю
1	ГОСТ 34.971–91 (ИСО 8822–88) Информационная технология (ИТ). Взаимосвязь открытых систем. Определение услуг уровня представления с установлением соединения	0,6944633	0,1744794
	ГОСТ Р ИСО/МЭК 9066–1–93 Системы обработки информации. Передача текста. Надежная передача. Часть 1. Модель и определение услуг		
2	ГОСТ 28147–89 Системы обработки информации. Защита криптографическая. Алгоритм криптографического преобразования	0,6806692	0,17162557
	ГОСТ Р 34.13–2015 Информационная технология (ИТ). Криптографическая защита информации. Режимы работы блочных шифров (с поправкой)		
3	ГОСТ Р 34.964–92 (ИСО 8602–87) Информационная технология (ИТ). Взаимосвязь открытых систем. Протокол транспортного уровня в режиме без установления соединения	0,7068537	0,1946876
	ГОСТ Р ИСО/МЭК 10025–3–94 Информационная технология (ИТ). Передача данных и обмен информацией между системами. Аттестационное тестирование транспортного уровня в режиме с установлением соединения при использовании услуг сетевого уровня в режиме с установлением соединения. Часть 3. Спецификация протокола административного управления тестированием		

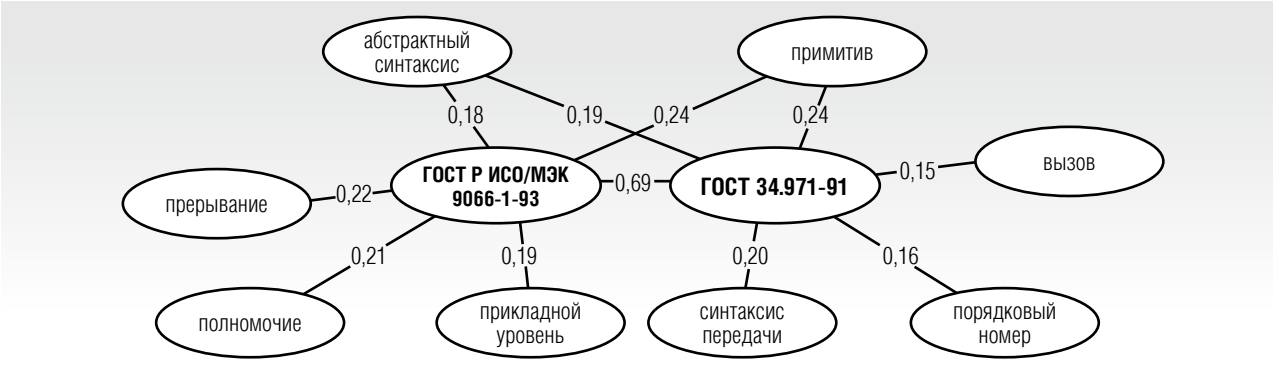

Рис. 3. Граф связей пары документов ГОСТ 34.971–91 и ГОСТ Р ИСО/МЭК 9066–1–93 и пяти наиболее значимых для каждого из них терминов

Пример 3. На *рисунке 3* представлен один из вариантов графа связей документов и терминов $\mathbf{G}_{TFIDF}$, а именно – его фрагмент с рассмотренными выше документами ГОСТ 34.971-91 и ГОСТ Р ИСО/МЭК 9066-1-93. Для каждого из документов показаны связи с пятью наиболее «весомыми» (значимыми) терминами.

Эксперимент был повторен на другом наборе документов: N_{D_RW} = 218 документов в формате docx, относящихся к тематике железных дорог. Для документов был сформирован массив полученных текстовых строк M_{RW} и выделен набор терминов T_{RW}.

Результаты второго эксперимента подтвердили сделанный вывод о том, что, несмотря на смещение кривой распределения числа пар по степени схожести, выявленная тенденция сохранена, а именно: при расчете близости методом TF-IDF с векторным разложением по словарю терминов в наборе обнаруживается больше схожих документов, чем при разложении по всему словарю корпуса.

Разница между матрицей схожести документов по терминам и матрицей схожести по общему словарю представлена на графиках *на рисунке 4.*

5. Дискуссия

Предложенный алгоритм дает существенное отличие при поиске близких документов в диапазоне схожести 60–80%. При более высокой степени схожести документов распределение близости по специализированным терминам не сильно отличается от распределения близости по полному словарю, поскольку концентрация терминов в документах приближается к концентрации общеупотребимых слов.

Из графика на *рисунке 2* видно, что для данного набора документов высокая близость по словарю означает высокую близость по терминам. Однако в обратном случае поиск документов по специализированным терминам дает более высокую степень схожести.

Приведенные в *таблице 3* примеры документов демонстрируют, что описанный метод расчета схожести по ранжирующей функции с разложением по словарю терминов способен снизить риск потери той части решения, где схожесть оказывается ниже

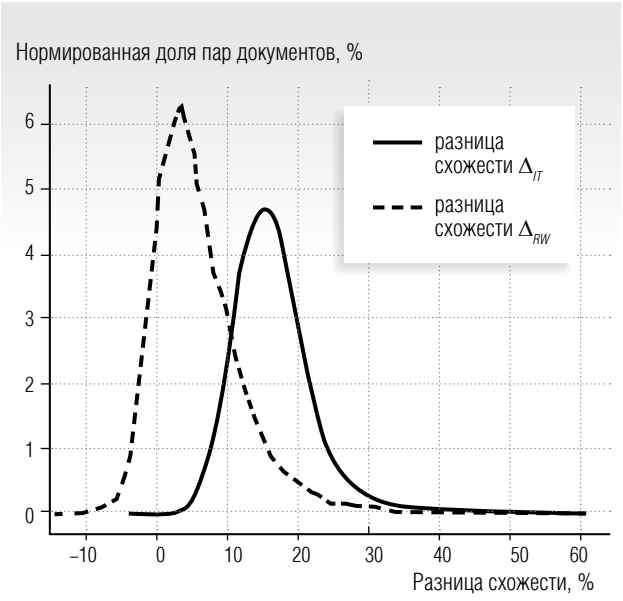

Рис. 4. Распределение разницы в степени схожести пар документов, вычисленной по двум словарям: все слова корпуса за вычетом редко встречающихся и словарь из специализированных терминов (Δ_{IT} – разница на корпусе M_{IT}, Δ_{RW} – разница на корпусе M_{RW})

установленного порога, но документы являются схожими с точки зрения пользователя системы.

Следует отметить, что в общем случае при переходе от общего словаря к терминологии при поиске схожести теоретически возможна потеря части решения – документов, схожих согласно человеческой оценке. Однако в реальных информационных системах эта потеря несущественна, так как пользователь в первую очередь сфокусирован на выявлении схожести в рамках определенной темы, то есть схожести на основании специфической для предметной области лексики. Поэтому при переходе от общего словаря к терминам возможна потеря лишь тех пар схожих документов, которые близки за счет общеупотребительной лексики. Количественные показатели таких потерь зависят от доли терминов в словаре, которая меняется от корпуса к корпусу. Для вычисления вклада общеупотребительных слов в общее значение степени схожести могут использоваться различные метрики, разработка которых выходит за рамки настоящего исследования.

В проведенном эксперименте не было обнаружено документов, у которых схожесть по общему словарю была бы больше, чем схожесть по терминам. На графике, демонстрирующем разницу между схожестью по словарю терминов и общему словарю, количество отрицательных значений незначительно, а разница между схожестью у таких пар документов составляет сотые доли процента.

Сравнение результатов двух экспериментов (на массивах M_{IT} и M_{RW}) позволило сделать дополнительные выводы и выделить факторы, влияющие на качество результата при поиске схожих документов с использованием векторного разложения по словарю терминов, в том числе созданному автоматически, на основе того же набора документов.

Неопределенность, вносимая в процесс обработки данных на шаге распознавания текста в документах, изначально полученных в графических форматах, является ключевым фактором при анализе корпуса документов. Недостаточное качество распознавания, когда часть текста теряется, снижает среднюю схожесть по TF-IDF при использовании общего словаря, однако по словарю терминов такая близость документов по-прежнему обнаруживается. Таким образом, предложенный алгоритм позволяет избежать затрат на подготовку текста для анализа – предварительную очистку текстового слоя и структурирование документа для поиска ключевых разделов.

Результат зависит и от выбора метода для автоматического составления словаря терминов. Эксперимент показал, что если термины выделяются из текста специального раздела анализируемых документов, то на результаты влияет наличие описанных терминов в содержательной части документа: если термины лишь перечислены в соответствующем разделе, но далее не используются, то значение схожести между документами искажается.

При автоматическом составлении словаря терминов схожесть также зависит от количества ошибок оформления в описании терминов в документах, поскольку в этом случае часть словаря может быть утеряна: термины используются в тексте документа, но не включены в словарь (базис для векторного разложения).

В общем случае схожесть документов по терминам зависит от выбора диапазона для *n*-грамм при составлении словаря терминов: длинные термины, состоящие из нескольких слов, в тексте документа с большой вероятностью ни разу не будут обнаружены целиком, когда все слова *n*-граммы входят в текст в нужном порядке и стоят подряд.

Пример 4. В документе из корпуса M_{RW} «ГОСТ 33798.4-2016 (IEC 60077-4:2003). Электрооборудование железнодорожного подвижного состава. Часть 4. Выключатели автоматические переменного тока. Общие технические условия» вводятся, в числе прочих, следующие термины:

✦ «Ручной рычаг включения/отключения: Рычаг для приведения автоматического выключателя во включенное или отключенное состояние ручным способом». Указанный термин ни разу не используется в содержательной части текста данного документа.

✦ «Выключатель внутренней установки: Выключатель, предназначенный только для установки и эксплуатации с защитой от неблагоприятных условий эксплуатации (ветра, дождя, снега, повышенного отложения грязи, нестандартных окружающих условий, льда и изморози)». В полном виде термин в тексте не обнаруживается, используется в другой форме только один раз: «Выключатели классифицируют: ... в соответствии с типом конструкции, то есть выключатели наружной или внутренней установки». Это означает, что данная триграмма не встречается в документе нигде, кроме самого описания термина. Таким образом, в данном случае TF (важность термина для данного документа) меняется. Поэтому при расчете схожести документов

по TF-IDF на основании словаря терминов целесообразно включать в него только униграммы и биграммы.

Наконец, если в структуре термина присутствует инверсия, то он также с высокой вероятностью не обнаруживается в тексте как исходная *n*-грамма.

Пример 5. В рассмотренном выше документе ГОСТ 33798.4-2016 определяется термин «выключатель полупроводниковый». В соответствии с нормами русского языка в связном тексте слова данного термина скорее всего поменяются местами: прилагательное разместится перед существительным. Действительно, в содержательной части документа указанный термин обнаруживается лишь один раз, без инверсии: «Для каждого выключателя указывают: вид устройства (например, воздушный выключатель, вакуумный выключатель, ..., полупроводниковый выключатель...».

Описанный эксперимент повторялся на корпусах одного типа. При проведении его на любом другом наборе документов результат может оказаться менее выразительным. Этому могут способствовать как указанные выше факторы, так и другие особенности документов: их структура, стилистика изложения, отраслевая специфика. Кроме того, в общем случае влияние на результат может оказать и тематическая сфокусированность корпуса, а также этимологические особенности терминологии, такие как использование общеупотребительной лексики в качестве узкоспециализированных терминов. Однако предварительно проведенные теоретические исследования позволяют сделать вывод о том, что при рассмотрении любого другого корпуса документов, в той или иной степени объединенных тематически, тенденция увеличения числа найденных схожих документов при разложении по словарю терминов сохранится.

Предложенный в работе подход реализован на практике – в рамках информационной системы Naumen LegalTech (продукт разрабатывается при грантовой поддержке Российского фонда развития информационных технологий). Это платформа аналитической обработки большого потока юридически значимых документов, ориентированная на пользователей, чьи интересы далеки от сферы информационных технологий и точных наук: юристов, методологов, разработчиков внутренней нормативной документации, авторов законодательных инициатив, специалистов по стандартизации, представителей отделов правовой экспертизы. Специфика работы целевой аудитории продукта такова, что в случае ошибки существует риск возникновения правовых последствий, но при этом интеллектуальные алгоритмы существенно помогают, работая в рекомендательном режиме. Для таких пользователей особенно важно доверять рекомендациям информационной системы, не погружаясь в технические детали ее реализации, а объяснимость решений является одним из ключевых факторов.

Заключение

В представленном исследовании рассмотрена возможность применения словаря специализированных терминов как частного случая тезауруса при векторном разложении корпуса документов с помощью ранжирующей функции.

Изучена проблема потери части решения в задаче поиска в корпусе текстовых документов пар элементов, близких по смыслу в значении «пословной» схожести текстов с учетом значимости термов.

В работе описан проведенный эксперимент на двух наборах нормативно-технических документов. Результаты эксперимента показали, что разложение по словарю терминов позволяет уменьшить потерю части решений, которая не удовлетворяет заданному условию превышения порогового значения степени схожести в паре документов, однако должна быть признана значимой согласно человеческой оценке.

Определены возможные факторы, влияющие на качество решения при выбранном методе поиска схожих документов в корпусе: выбор способа построения словаря терминов, ограничение длины *n*-граммы в словаре, грамматические особенности терминов.

Модели корпуса документов, основанные на специализированном терминологическом словаре, применимы в разных ИТ-решениях, учитывающих схожесть текстов: в рекомендательных системах, в модулях семантического поиска, при маршрутизации обращений. Предложенный алгоритм повышает точность и объяснимость рекомендаций в информационных системах, что ускоряет принятие решений пользователем и увеличивает эффективность его работы с большим объемом текстовых документов. ■

Литература

1. Краснов Ф.В., Смазневич И.С. Фактор объяснимости алгоритма в задачах поиска схожести текстовых документов // Вычислительные технологии. 2020. Т. 25. № 5. С. 107–123. DOI: 10.25743/ICT.2020.25.5.009.
2. Отраднов К.К., Жуков Д.О., Новикова О.А. Модель кластеризации слабоструктурированных текстовых данных // Современные информационные технологии и ИТ-образование. 2017. Т. 13. № 3. [Электронный ресурс]: http://sitito.cs.msu.ru/index.php/SITITO/article/view/295. DOI: 10.25559/SITITO.2017.3.439.
3. Calzolari N. An empirical approach to circularity in dictionary definitions // Cahiers de Lexicologie. 1977. Т. 31. №. 2. С. 118–128.
4. Amsler R.A. The structure of the Merriam-Webster pocket dictionary. Austin, TX: The University of Texas, 1980.
5. Grefenstette G. Explorations in automatic thesaurus discovery. New York: Springer Science & Business Media, 1994. DOI: 10.1007/978-1-4615-2710-7.
6. Bullinaria J.A. Semantic categorization using simple word co-occurrence statistics // ESSLLI Workshop on Distributional Lexical Semantics (ESSLLI 2008). Hamburg, Germany, 4–9 August 2008. P. 1–8.
7. Nazar R., Vivaldi J., Wanner L. Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora // Procesamiento del Lenguaje Natural. 2012. No 49. P. 67–74.
8. Chasing hypernyms in vector spaces with entropy / E. Santus [et al.] // 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 26–30 April 2014. P. 38–42.
9. Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 1972. Vol. 28. No 1. P. 11–21.
10. Jones K.S. Index term weighting // Information Storage and Retrieval. 1973. Vol. 9. No 11. P. 619–633.
11. Salton G., Yang C.S. On the specification of term values in automatic indexing. Ithaca, NY: Cornell University, 1973.
12. Salton G., Wong A., Yang C.S. A vector space model for automatic indexing // Communications of the ACM. 1975. Vol. 18. No 11. P. 613–620.
13. Отраднов К.К., Раев В.К. Экспериментальное исследование эффективности методик векторизации текстовых документов и алгоритмов их кластеризации // Вестник Рязанского государственного радиотехнического университета. 2018. № 64. С. 73–84. DOI: 10.21667/1995-4565-2018-64-2-73-84.
14. Bafna P., Pramod D., Vaidya A. Document clustering: TF-IDF approach // 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT 2016). Chennai, India, 3–5 Mach 2016. P. 61–66. DOI: 10.1109/ICEEOT.2016.7754750.
15. Fomin S.A., Belousov R.L. Detecting semantic duplicates in short news items // Business Informatics. 2017. No 2. P. 47–56. DOI: 10.17323/1998-0663.2017.2.47.56.
16. Qaiser S., Ali R. Text mining: Use of TF-IDF to examine the relevance of words to documents // International Journal of Computer Applications. 2018. Vol. 181. No 1. P. 25–29. DOI: 10.5120/ijca2018917395.
17. Krasnov F., Ushmaev O. Exploration of hidden research directions in oil and gas industry via full text analysis of OnePetro digital library // International Journal of Open Information Technologies. 2018. Vol. 6. No 5. P. 7–14.
18. Kim S.-W., Gil J.-M. Research paper classification systems based on TF-IDF and LDA schemes // Human-Centric Computing and Information Sciences. 2019. No 9. Article no 30. DOI: 10.1186/s13673-019-0192-7.
19. Hersh W.R., Monarch I.A., Lefferts R.G., Handerson S.K. Automatic indexing of abstracts via natural-language processing using a simple thesaurus / D.A. Evans [et al.] // Medical Decision Making. 1991. Vol. 11. No 4 (suppl.). P. 108–115.
20. Medelyan O., Witten I.H. Thesaurus based automatic keyphrase indexing // 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06). Chapel Hill, NC, USA, 11–15 June 2006. P. 296–297. DOI: 10.1145/1141753.1141819.
21. Голицына О.Л., Максимов Н.В., Федорова В.А. К определению семантической близости на основе связей объединенного тезауруса // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2016. №. 6. С. 30–44.
22. Loukachevitch N., Nokel M., Ivanov K. Combining thesaurus knowledge and probabilistic topic models // 6th International Conference on Analysis of Images, Social Networks, and Texts (AIST 2017). Moscow, Russia, 27–29 July 2017. P. 59–71.
23. Andrzejewski D., Zhu X., Craven M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors // 26th Annual International Conference on Machine Learning (ICML 2009). Montreal, Canada, 14–18 June 2009. P. 25–32. DOI: 10.1145/1553374.1553378.
24. Loukachevitch N., Ivanov K. Evaluating thesaurus-based topic models // 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018). Paris, France, 13–15 June 2018. P. 364–376. DOI: 10.1007/978-3-319-91947-8_38.
25. ГОСТ Р ИСО 704-2010. Терминологическая работа. Принципы и методы. М.: Стандартинформ, 2012.
26. Бунин М.С., Пирумова Л.Н. Информационно-поисковый тезаурус по сельскому хозяйству и продовольствию Центральной научной сельскохозяйственной библиотеки // Российская сельскохозяйственная наука. 2020. № 5. С. 72–75. DOI: 10.31857/S2500262720050178.
27. Aubin S., Hamon T. Improving term extraction with terminological resources // 5th International Conference on Natural Language Processing (FinTAL 2006). Turku, Finland. 23–25 August 2006. P. 380–387.

Об авторах

Краснов Федор Владимирович
кандидат технических наук;
эксперт департамента информационных систем управления, NAUMEN R&D, 620028, г. Екатеринбург, ул. Татищева, д. 49А;
E-mail: fkrasnov@naumen.ru
ORCID: 0000-0002-9881-7371

Смазневич Ирина Сергеевна
бизнес-аналитик департамента информационных систем управления, NAUMEN R&D, 620028, г. Екатеринбург, ул. Татищева, д. 49А;
E-mail: ismaznevich@naumen.ru
ORCID: 0000-0002-5996-4635

Баскакова Елена Николаевна
ведущий системный аналитик департамента информационных систем управления, NAUMEN R&D, 620028, г. Екатеринбург, ул. Татищева, д. 49А;
E-mail: enbaskakova@naumen.ru
ORCID: 0000-0002-7071-8961

The problem of loss of solutions in the task of searching similar documents: Applying terminology in the construction of a corpus vector model

Fedor V. Krasnov
E-mail: fkrasnov@naumen.ru

Irina S. Smaznevich
E-mail: ismaznevich@naumen.ru

Elena N. Baskakova
E-mail: enbaskakova@naumen.ru

NAUMEN R&D
Address: 49A, Tatishcheva Street, Ekaterinburg 620028, Russia

Abstract

This article considers the problem of finding text documents similar in meaning in the corpus. We investigate a problem arising when developing applied intelligent information systems that is non-detection of a part of solutions by the TF-IDF algorithm: one can lose some document pairs that are similar according to human assessment, but receive a low similarity assessment from the program. A modification of the algorithm, with the replacement of the complete vocabulary with a vocabulary of specific terms is proposed. The addition of thesauri when building a corpus vector model based on a ranking function has not been previously investigated; the use of thesauri has so far been studied only to improve topic models. The purpose of this work is to improve the quality of the solution by minimizing the loss of its significant part and not adding "false similar" pairs of documents. The improvement is provided by the use of a vocabulary of specific terms extracted from the text of the analyzed documents when calculating the TF-IDF values for corpus vector representation. The experiment was carried out on two corpora of structured normative and technical documents united by a subject: state standards related to information technology and to the field of railways. The glossary of specific terms was compiled by automatic analysis of the text of the documents under consideration,

and rule-based NER methods were used. It was demonstrated that the calculation of TF-IDF based on the terminology vocabulary gives more relevant results for the problem under study, which confirmed the hypothesis put forward. The proposed method is less dependent on the shortcomings of the text layer (such as recognition errors) than the calculation of the documents' proximity using the complete vocabulary of the corpus. We determined the factors that can affect the quality of the decision: the way of compiling a terminology vocabulary, the choice of the range of *n*-grams for the vocabulary, the correctness of the wording of specific terms and the validity of their inclusion in the glossary of the document. The findings can be used to solve applied problems related to the search for documents that are close in meaning, such as semantic search, taking into account the subject area, corporate search in multi-user mode, detection of hidden plagiarism, identification of contradictions in a collection of documents, determination of novelty in documents when building a knowledge base.

Key words: similarity of documents; semantic proximity; thesauri application; corpus vector model; applied intelligent information systems; algorithm explainability; similarity evaluation; text mining.

Citation: Krasnov F.V., Smaznevich I.S., Baskakova E.N. (2021) The problem of loss of solutions in the task of searching similar documents: Applying terminology in the construction of a corpus vector model. *Business Informatics*, vol. 15, no 2, pp. 60–74. DOI: 10.17323/2587-814X.2021.2.60.74

References

1. Krasnov F.V., Smaznevich I.S. (2020) The explicability factor of the algorithm in the problems of searching for the similarity of text documents. *Computational Technologies*, vol. 25, no 5, pp. 107–123 (in Russian). DOI: 10.25743/ICT.2020.25.5.009.
2. Otradnov K.K., Zhukov D.O., Novikova O.A. (2017) Clustering model of low-structured text data. *Modern Information Technologies and IT-Education*, vol. 13, no 3. Available at: http://sitito.cs.msu.ru/index.php/SITITO/article/view/295 (in Russian). DOI: 10.25559/SITITO.2017.3.439.
3. Calzolari N. (1977) An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie*, vol. 31, no 2, pp. 118–128.
4. Amsler R.A. (1980) *The structure of the Merriam-Webster pocket dictionary*. Austin, TX: The University of Texas.
5. Grefenstette G. (1994) *Explorations in automatic thesaurus discovery*. New York: Springer Science & Business Media. DOI: 10.1007/978-1-4615-2710-7.
6. Bullinaria J.A. (2008) Semantic categorization using simple word co-occurrence statistics. Proceedings of the *ESSLLI Workshop on Distributional Lexical Semantics (ESSLLI 2008), Hamburg, Germany, 4–9 August 2008*, pp. 1–8.
7. Nazar R., Vivaldi J., Wanner L. (2012) Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del Lenguaje Natural*, no 49, pp. 67–74.
8. Santus E., Lenci A., Lu Q., Walde S.S. (2014) Chasing hypernyms in vector spaces with entropy. Proceedings of the *14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014*, pp. 38–42.
9. Jones K.S. (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, vol. 28, no 1, pp. 11–21.
10. Jones K.S. (1973) Index term weighting. *Information Storage and Retrieval*, vol. 9, no 11, pp. 619–633.
11. Salton G., Yang C.S. (1973) *On the specification of term values in automatic indexing*. Ithaca, NY: Cornell University.
12. Salton G., Wong A., Yang C.S. (1975) A vector space model for automatic indexing. *Communications of the ACM*, vol. 18, no 11, pp. 613–620.
13. Otradnov K.K., Raev V.K. (2018) Experimental study of text documents vectorization techniques and their clustering algorithms efficiency. *Vestnik of Ryazan State Radio Engineering University*, no 64, pp. 73–84 (in Russian). DOI: 10.21667/1995-4565-2018-64-2-73-84.
14. Bafna P., Pramod D., Vaidya A. (2016) Document clustering: TF-IDF approach. Proceedings of the *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT 2016), Chennai, India, 3–5 Mach 2016*, pp. 61–66. DOI: 10.1109/ICEEOT.2016.7754750.
15. Fomin S.A., Belousov R.L. (2017) Detecting semantic duplicates in short news items. *Business Informatics*, no 2, pp. 47–56. DOI: 10.17323/1998-0663.2017.2.47.56.
16. Qaiser S., Ali R. (2018) Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, vol. 181, no 1, pp. 25–29. DOI: 10.5120/ijca2018917395.
17. Krasnov F., Ushmaev O. (2018) Exploration of hidden research directions in oil and gas industry via full text analysis of OnePetro digital library. *International Journal of Open Information Technologies*, vol. 6, no 5, pp. 7–14.
18. Kim S.-W., Gil J.-M. (2019) Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, no 9, article no 30. DOI: 10.1186/s13673-019-0192-7.
19. Evans D.A., Hersh W.R., Monarch I.A., Lefferts R.G., Handerson S.K. (1991) Automatic indexing of abstracts via natural-language processing using a simple thesaurus. *Medical Decision Making*, vol. 11, no 4 (suppl.), pp. 108–115.
20. Medelyan O., Witten I.H. (2006) Thesaurus based automatic keyphrase indexing. Proceedings of the *6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06), Chapel Hill, NC, USA, 11–15 June 2006*, pp. 296–297. DOI: 10.1145/1141753.1141819.
21. Golitsyna O.L., Maksimov N.V., Fyodorova V.A. (2016) On the definition of semantic proximity based on the links of the combined thesaurus. *Scientific and Technical Information. Series 2: Information Processes and Systems*, no 6, pp. 30–44 (in Russian).

22. Loukachevitch N., Nokel M., Ivanov K. (2017) Combining thesaurus knowledge and probabilistic topic models. Proceedings of the *6th International Conference on Analysis of Images, Social Networks, and Texts (AIST 2017), Moscow, Russia, 27–29 July 2017*, pp. 59–71.

23. Andrzejewski D., Zhu X., Craven M. (2009) Incorporating domain knowledge into topic modeling via Dirichlet forest priors. Proceedings of the *26th Annual International Conference on Machine Learning (ICML 2009), Montreal, Canada, 14–18 June 2009*, pp. 25–32. DOI: 10.1145/1553374.1553378.

24. Loukachevitch N., Ivanov K. (2018) Evaluating thesaurus-based topic models. Proceedings of the *23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018), Paris, France, 13–15 June 2018.* P. 364–376. DOI: 10.1007/978-3-319-91947-8_38.

25. Standardinform (2012) *GOST R ISO 704-2010. Terminology work. Principles and methods.* Moscow: Standardinform (in Russian).

26. Bunin M.S., Pirumova L.N. (2020) Information and search thesaurus on agriculture and food of the Central Scientific Agricultural Library, *Russian Agricultural Science*, no 5, pp. 72–75 (in Russian). DOI: 10.31857/S2500262720050178.

27. Aubin S., Hamon T. (2006) Improving term extraction with terminological resources. Proceedings of the *5th International Conference on Natural Language Processing (FinTAL 2006), Turku, Finland, 23–25 August 2006*, p. 380–387.

About the authors

Fedor V. Krasnov

Cand. Sci. (Tech.);

Expert, Department of Management Information Systems, NAUMEN R&D, 620028, 49A, Tatishcheva Street, Ekaterinburg 620028, Russia;

E-mail: fkrasnov@naumen.ru

ORCID: 0000-0002-9881-7371

Irina S. Smaznevich

Business Analyst, Department of Management Information Systems, NAUMEN R&D, 620028, 49A, Tatishcheva Street, Ekaterinburg 620028, Russia;

E-mail ismaznevich@naumen.ru

ORCID: 0000-0002-5996-4635

Elena N. Baskakova

Leading System Analyst, Department of Management Information Systems, NAUMEN R&D, 620028, 49A, Tatishcheva Street, Ekaterinburg 620028, Russia;

E-mail enbaskakova@naumen.ru

ORCID: 0000-0002-7071-8961