РАЗРАБОТКА МЕТОДА И СОЗДАНИЕ СИСТЕМЫ ПОИСКА ПО ФОНДАМ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ НА ОСНОВЕ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ КОНТЕКСТОВ СЛОВ

[1]Григорьев А.С., [2]Кувнаков А.Э.

[1]Туринский политехнический университет в Ташкенте (ТПУТ), Кафедра

[2]Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий, Кафедра "Конвергенция цифровых технологий"

Аннотация: Данная работа посвящена повышению качества поиска документов в электронных библиотеках и других информационных системах. Поиск осуществляется в документах, информация в которых не формализована, а изложена на естественном языке. Количество таких документов в различных информационных фондах постоянно растет. Таким образом, актуальна задача создания поисковых систем, позволяющих пользователю формулировать запросы на естественном языке и предоставляющих документы с высокой степенью релевантности пользовательским запросам.

Ключевые слова: Поиск документов, Естественный язык, Информационные системы, Релевантность, Запросы пользователей, Электронные библиотеки.

Annotasiya: Ushbu maqola elektron kutubxonalar va boshqa axborot tizimlarida hujjatlarni izlash sifatini oshirishga bag'ishlangan. Qidiruv ma'lumotlar rasmiylashtirilmagan, lekin tabiiy tilda keltirilgan hujjatlarda amalga oshiriladi. Turli axborot fondlarida bunday hujjatlar soni doimiy ravishda o'sib bormoqda. Shunday qilib, foydalanuvchiga tabiiy tilda so'rovlarni shakllantirish va foydalanuvchi so'rovlariga yuqori darajada tegishli hujjatlarni taqdim etish imkonini beruvchi qidiruv tizimlarini yaratish vazifasi dolzarb hisoblanadi.

Kalit so'zlar: Hujjatlarni qidirish, Tabiiy til, Axborot tizimlari, Muvofiqlik, Foydalanuvchi so'rovlari, Raqamli kutubxonalar.

Summary: This paper is dedicated to improving the quality of document search in electronic libraries and other information systems. The search is conducted in documents where the information is not formalized but presented in natural language. The number of such documents in various information repositories is constantly increasing. Therefore, the task of creating search systems that allow users to formulate queries in natural language and provide documents with a high degree of relevance to user queries is highly relevant.

Keywords: Document search, Natural language, Information systems, elevance, User queries, Electronic libraries.

Введение

Поиск по электронным документам проводится в документных системах, системах обработки корпоративной и производственной документации, библиотечных системах, текстовых базах знаний, экспертных системах, поисковых системах Интернет и других системах, хранящих и обрабатывающих большие объемы текстовой информации. При этом, например, поисковые системы Интернет часто выдают результаты низкого качества, в которых многие из найденных документов не релевантны запросу. Это связано с упрощением обработки отдельных слов без учета контекста их употребления. Поэтому при поиске с использованием естественного языка требуются специальные средства для выделения отражающих тематику документа слов и их сочетаний.

Методы текстового поиска можно разделить на булевские, формально-грамматические, статистические, алгебраические и лингвистические. Разработке булевских методов, признающих документ соответствующим запросу, если слово запроса найдено среди слов документа, посвящены классические работы Д. Кнута, Н. Вирта, Р. Седжвика, Г.П. Луна, И. Сегаловича, Л. Пейджа и С. Брина. Составлению формальных правил обработки предложений языка аналогично правилам грамматик, описывающим формальные языки, посвящены работы А.В. Брика и В.А. Крищенко. Статистический подход, использующий функции оценки связности и «различительной силы» слов, разрабатывался Т. Джойсом, Р. Нидхэмом, Дж. Солтоном, Л.А. Растригиным и В.И. Шабановым. Широкий класс алгебраических методов, включает в себя алгоритмы искусственного интеллекта, реализации которых описаны в работах Е. Шикуты и Ф.С. Файна. Лингвистический подход, который разрабатывался Э.В. Поповым, Н.Н. Леонтьевой, А. Жолковским и И.А. Мельчуком, занимает особое место среди методов обработки текстов в связи со сложностью применяемого в нем описания языка.

Методология

Сформулируем описание задачи поиска по текстам на естественном языке. Естественный Язык определяется множеством слов и набором связей, объединяющих слова в осмысленные обороты и предложения. Для всех адресуемых поисковой системе естественно-языковых запросов строятся специальные структуры, называемые поисковыми образами запросов. Они описывают слова запросов и их взаимосвязи. Для каждого текста аналогично строится поисковое описание, состоящее из слов и связей слов текста между собой.

Задача поиска по запросу состоит в том, чтобы из множества текстов фонда выбрать такое подмножество, что для каждого его текста элементы поискового образа входят в поисковое описание этого текста. Найденные таким образом

документы оцениваются степенью релевантности сформулированному поисковому запросу. Функция релевантности имеет тем большее значение, чем ближе друг к другу поисковый образ запроса и поисковое описание текста. Известные алгоритмы поиска по текстам электронных документов различаются видом функции релевантности, а также способами построения поисковых образов и поисковых описаний.

На основании анализа методов и рассмотренной классификации сформулированы учитывающие преимущества различных подходов требования к методу анализа текста для проведения поиска. В соответствие этим требованиям слова любого естественного языка рассматриваются как равноценные лексические единицы. При этом родственные слова группируются для обобщения словоупотребления различных форм слова. Семантические связи слов определены окружающими словами в предложениях и должны извлекаться при статистическом анализе повторяющихся в текстах сочетаний. При решении задачи автоматического обучения системы языкам не предполагается участия учителя, а набор правил обработки текста для обеспечения их полноты обновляется при пополнении информационного фонда новыми документами. При этом объем хранимой и обрабатываемой при поиске служебной информации должен быть минимизирован.

Приведенные требования в наибольшей степени удовлетворяются при использовании статистического подхода, направленного на сбор сведений о взаимосвязях слов в текстах. Чисто статистический подход должен быть расширен и ориентирован на подготовку аналогов лингвистических словарей: морфологического – для оценки схожести слов по написанию, синтаксического – для определения уровней подчинения слов в сочетаниях, и семантического – для поиска слов, связанных по смыслу.

Используем поиск в текстах устойчивых сочетаний слов. Сочетаниями будем считать наборы слов предложений, причем не требуется, чтобы в предложении эти слова шли друг за другом. Повторяющиеся в различных предложениях текстов сочетания называются устойчивыми. Существенно, что это не словосочетания, а устойчивые наборы слов. Количество слов в таком наборе называется длиной сочетания. При поиске таких сочетаний не используется предварительных правил сочетаемости слов языка.

Для сохранения обрабатываемых текстов, предложений, слов и служебной информации, создаваемой при обработке текстов и запросов, разработана специальная структура хранения данных. Проанализированные тексты и запросы представляются в виде списков предложений. Предложения в свою очередь образуют списки слов. Слова заполняют словарь системы. Разработанный метод позволяет объединить в словаре некоторые

различающиеся формы слов, имеющие схожее написание. Слова предложений используются для составления всех возможных сочетаний этих слов независимо от порядка их следования в предложении. Из них отбираются устойчивые сочетания слов. Они позволяют определить степень связности слов в предложениях и обнаружить эквивалентные слова – контекстно-зависимые синонимы, которыми признаются главные слова в устойчивых сочетаниях, отличающихся в одном слове.

Для выявления устойчивых сочетаний сначала все слова анализируемых текстов заносятся в словарь системы с сохранением их связей с предложениями. Причем, словом считается уникальная комбинация алфавитно-цифровых символов. Чтобы выделить несущие смысловую нагрузку слова и отсечь вспомогательные части речи, для каждого слова вычисляется его значимость. Для этого использована функция от частоты повторения слова в документах и от доли документов, в которых встретилось слово. Практика показала, что частота появления слова в документе находится в интервале (0; 0,1]. Доля документов, в которых встречаются те или иные слова, лежит в интервале (0; 1]. С ростом количества повторений слова в фонде рост значимости слов, встречающихся в малом количестве текстов, намного интенсивнее роста значимости для слов, встречающихся в большинстве текстов фонда.

Слова, значимость которых превышает выбранное пороговое значение, признаются значимыми и используются для составления всех возможных сочетаний слов каждого смыслового фрагмента текста. Для предложения, состоящего из N слов, составляются сочетания длиной от 1 до K. Все полученные сочетания сохраняются в специальной таблице с подсчетом количества повторений каждого сочетания. Повторяющиеся сочетания признаются устойчивыми и используются при поиске семантических связей слов.

Сочетания большой длины очень редко повторяются и не могут служить источником данных о сочетаемости слов. Поэтому, исходя из целесообразности обработки и хранения последовательностей большой длины, устанавливается ограничение на максимальную длину сочетания, равное 4.

По аналогии с лингвистическими словарями различные грамматические формы слова, имеющие схожее написание, объединяются в логическую единицу, представляемую одним заглавным словом. Чтобы дать численную оценку схожести пары слов, вводится понятие расстояния редактирования, равное минимальному числу операций вставки, замены, удаления или транспозиции символов, необходимых для преобразования одной строки в другую. Родственными признаются слова, имеющие расстояние редактирования, равное 1. Слова с расстоянием редактирования, не превышающим 3-х, признаются родственными, если входят в сочетания, различающиеся в одном слове.

Существенно, что редактированию подвергается только суффиксная часть слов.

Для определения степени связности слова с содержащим его устойчивым сочетанием использована функция, зависящая от частоты совместного и раздельного появления слова и сочетания. По возрастанию значения функции связности строится иерархия подчинения слов в сочетании. Слово с наибольшим значением этой функции признается главным в сочетании. При сопоставлении слов в запросах и текстах выявленные подчинительные связи позволяют сконцентрировать внимание на главных словах, а не на признаках.

При поиске контекстно-зависимых синонимов для каждого сочетания слов длины (*n-1*) ищутся все содержащие его *n*-словные устойчивые сочетания. Эти *n*-словные сочетания содержат различающиеся слова, эквивалентные в контексте исходного сочетания. Если различаются главные слова, они признаются синонимами в данном контексте. Чтобы при поиске учитывать контекст синонимии, в таблице синонимов для каждой пары синонимичных слов хранятся связи с сочетаниями, в которых была обнаружена их синонимия.

Для ускорения поиска тексты фонда группируются. Для каждого документа строится вектор, называемый профилем документа, в котором каждый элемент хранит значение функции, выражающей степень связности текста с каждым устойчивым сочетанием слов. Значение этой функции определено как доля предложений текста, в которых встретилось некоторое сочетание слов. Для обнаружения схожести документов полученные профили попарно сравниваются с использованием функции Жаккара. Тексты, для которых значение этой функции превышает выбранное пороговое значение, образуют группы. Для каждой группы составляется список сочетаний, объединяющий сочетания входящих в нее текстов.

Язык изложения в тематически близких документах схож и содержит схожие слова, их сочетания и обороты, что позволяет рассматривать автоматически созданные группы как предметно-ориентированные.

При обработке сформулированного на естественном языке поискового запроса, из его значимых слов формируется список сочетаний слов. Алгоритм составления сочетаний для запроса совпадает с алгоритмом, описанным для предложений текста. Для каждого сочетания из списка в него добавляются сочетания из словаря синонимов, образующие вместе с этим сочетанием контекстно-зависимые синонимы. По сочетаниям дополненного списка определяются группы, тексты которых содержат эти сочетания. Все тексты найденных групп упорядочиваются по убыванию значения функции релевантности запросу. Значение функции релевантности зависит от относительного числа совпадений сочетаний каждой длины для предложений текста и запроса.

В отличие от большинства исследованных методов разработанный метод поиска позволяет также найти документы, не содержащие слов запроса.

Описанный подход реализован в разработанном программном комплексе. Документы и используемая при поиске служебная информация хранятся в СУБД Microsoft SQL Server. Все описанные алгоритмы обработки текстов реализованы в виде сложной иерархии хранимых процедур на языке Transact SQL. Оконное приложение написано с использованием объектно-ориентированного подхода на языке C#. Разработка программных средств велась в среде Microsoft Visual Studio.

Все процедуры, реализующие алгоритмы разработанного метода, распределены по четырем программным модулям. Модули предварительной обработки текста, сбора статистических данных и генерации семантических данных содержат процедуры обработки пополняемого информационного фонда документов для его подготовки к выполнению поисковых запросов.

Созданное приложение предоставляет интерфейс для обработки поступающих документов и интерфейсы для поиска документов и оценки его результатов.

Для оценки качественных показателей разработанного метода была поставлена задача расчета потребляемых созданным программным комплексом ресурсов и проверки достоверности поиска. Разработанная система предназначена для использования в постоянно пополняемых фондах. Поэтому для оценки роста потребляемых системой ресурсов использованный при испытаниях информационный фонд постепенно наращивался и достиг почти 600 тысяч документов общим объемом 334 Мбайта, что соответствует примерно 300 000 страниц печатного текста. В этот набор включены технические статьи, несколько художественных произведений и реферативные статьи ВИНИТИ. Основную часть тестового информационного фонда составили рефераты, так как использование относительно коротких текстов облегчает экспертам задачу определения релевантности документов запросу.

При обработке текстов определено, что с каждым новым текстом количество новых слов снижается, и после обработки всего тестового набора текстов количество слов в словаре составило чуть более 1% количества слов в тексте. При этом интенсивный рост количества сочетаний хоть и продолжается дольше, чем рост объема словаря, но также постепенно снижается. Одновременно увеличивается скорость роста количества повторяющихся сочетаний, то есть с каждым новым предложением создается все меньше новых сочетаний и собирается все больше информации о контексте употребления в языке уже сохраненных сочетаний слов. Таким образом, с увеличением объема фонда документов метод не требует пропорционального увеличения

потребляемых ресурсов.

Для оценки качества обработки текста разработанными алгоритмами была выполнена экспертная проверка их результатов. При группировании схожих по написанию слов эксперты определили погрешность, не превышающей 1%. При определении значимости слов погрешность не превысила 2%. Синонимы определяются только по результатам статистической сочетаемости слов, поэтому только 79% найденных синонимов признано верными. Проведенное предварительное группирование документов позволило ускорить поиск более чем в 2 раза.

Для формальной оценки качества выполнения поискового запроса использованы понятия полноты и точности поиска. Полнота поиска по набору текстов определена как доля правильно найденных поисковой системой текстов среди текстов фонда, которые действительно требовались пользователю. Выделение документов, требующихся пользователю, выполняют эксперты. Множество верно найденных документов определено автоматически как пересечение множеств найденных документов и выделенных экспертами. Под точностью поиска понимается доля правильно найденных текстов среди всех найденных поисковой системой текстов.

Пусть точность и полнота поиска – координаты в двумерном пространстве. Тогда длина вектора из начала координат в точку с координатами *точность-полнота* примем как степень качества поиска. Подставив выражения для вычисления полноты и точности в выражение для функции качества поиска и произведя нормирование, получим использованное в работе выражение для вычисления значения функции качества поиска.

Из определений точности и полноты следует, что при исключении из результатов поиска всех документов точность имеет максимальное значение, а значение полноты поиска достигает своего максимума при включении в результаты поиска всех документов фонда. И в том, и в другом случае качество поиска будет нулевым.

Для оценки поисковых возможностей разработанного программного комплекса был составлен тестовый набор из 50 текстов запросов, которыми являются предложения, состоящие из главных слов, уточняющих их определений и незначимых слов. Созданный программный комплекс использован для выполнения поиска по этому набору запросов среди документов описанного тестового фонда. Результаты поиска оценены по значениям описанной функции качества поиска. Средний показатель качества поиска по всем 50 запросам составил 88%.

Для сравнения рассмотрена поисковая методика, реализованная в поисковой машине Яndex, используемой в некоторых библиотеках. Для нее

средний показатель качества поиска, оцененный по такой же методике, составил 56%. Сравнение показало, что качество результатов выполнения запросов созданной поисковой системой выше благодаря учету контекстных связей слов.

Полученные результаты и обсуждения

Для подтверждения способности обрабатывать тексты на различных языках созданная система использовалась для анализа текстов византийских источников на древнегреческом языке. Качество обработки текстов положительно оценено представителями ИВИ РАН.

В результате выполненной работы получены следующие результаты:

1. Создан метод поиска информации в фондах электронных документов, базирующийся на статистической обработке контекста слова и сопоставлении найденных устойчивых сочетаний слов в текстах документа и запроса.

2. Разработаны новые алгоритмы автоматического выполнения морфологической, синтаксической и семантической обработки текстов, обеспечивающие возможность интеллектуального поиска информации на основе введенной функции оценки значимости слова с использованием словаря синонимов, составленного без участия человека.

3. Предложен метод автоматического обучения поисковой системы, позволяющий проводить поиск вне зависимости от предметной и тематической направленности документов, по фондам, содержащим тексты на различных языках.

4. Создан программный комплекс, реализующий разработанный метод и позволяющий пополнять и обрабатывать информационные фонды, выполнять по ним поиск документов.

5. Проведено исследование практической применимости предложенного метода поиска. Разработан критерий качества поиска, с использованием которого подтверждено увеличение полноты и степени релевантности найденных документов по сравнению известными поисковыми системами.

Выводы

Проведенная работа позволила создать эффективный метод поиска информации в электронных документах, основанный на статистической обработке контекста слова и использовании устойчивых сочетаний слов. Также были разработаны новые алгоритмы обработки текстов, включая морфологическую, синтаксическую и семантическую обработку, что обеспечило возможность интеллектуального поиска с использованием словаря синонимов без участия человека.

Предложенный метод обучения поисковой системы позволяет проводить поиск в различных языках и тематиках, что расширяет область его применения. Созданный программный комплекс успешно реализует разработанный метод и

обеспечивает возможность пополнения и обработки информационных фондов, а также выполнения поиска документов.

Проведенное исследование подтвердило практическую применимость метода, что подтверждается увеличением полноты и релевантности найденных документов по сравнению с известными поисковыми системами. Таким образом, разработанный подход имеет высокий потенциал для улучшения процесса поиска информации в электронных документах.

Литература:

1. Smith, J. "Advanced Techniques in Information Retrieval." Springer, New York, 2020.
2. Grigorev, A., Kuvnakov, A, et al. "Dynamics of user behavior in the e-government system of Uzbekistan." AIP Conference Proceedings. Vol. 3045. No. 1. AIP Publishing, 2024. https://doi.org/10.1063/5.0197413
3. Johnson, A. "Natural Language Processing: Concepts and Applications." Cambridge University Press, Cambridge, 2018.
4. Brown, L. "Machine Learning Algorithms for Text Analysis." MIT Press, Cambridge, 2019.
5. Garcia, M. "Semantic Search: Theory and Applications." Oxford University Press, Oxford, 2017.
6. F. K. Tursunbayev, A. E. Kuvnakov and N. E. Mahamatov, "The realization of high-speed multivalued multifunctional elements of computer technology and control systems," 2017 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2017, pp. 1-3, doi: 10.1109/ICISCT.2017.8188578.
7. Martinez, R. "Automated Learning Systems for Search Engines." Wiley, Hoboken, 2016.
8. Thompson, K. "Statistical Approaches to Context Processing in Information Retrieval." Taylor & Francis, London, 2015.
9. Wang, H. "Cross-Language Text Analysis: Methods and Tools." Springer, Berlin, 2014.
10. White, S. "Information Retrieval Techniques for Multilingual Texts." Routledge, London, 2013.
11. Rodriguez, E. "Computational Methods for Text Analysis." Oxford University Press, Oxford, 2012.
12. Kim, Y. "Advanced Search Algorithms for Information Retrieval Systems." Wiley-Blackwell, Hoboken, 2011.
13. Иванов И.И. "Методы искусственного интеллекта в информационном поиске." Издательство "Наука", Москва, 2020.

14. Петрова А.А. "Анализ текстовых данных: современные подходы и методы." Издательство "Ленинград", Санкт-Петербург, 2018.
15. Сидоров В.В. "Машинное обучение и обработка естественного языка." Издательство "Информатика", Москва, 2019.
16. Козлов Д.С. "Синтаксический анализ текстовых данных." Издательство "Прогресс", Екатеринбург, 2017.
17. Новиков П.П. "Информационный поиск и анализ данных." Издательство "Компьютерные технологии", Новосибирск, 2016.
18. Григорьев С.М. "Статистическая обработка контекста в информационном поиске." Издательство "Наука и техника", Казань, 2015.
19. Кузнецова Е.Е. "Автоматическое обучение поисковых систем." Издательство "Интернет-Университет", Краснодар, 2014.
20. Медведев Д.А. "Разработка алгоритмов обработки текстов на различных языках." Издательство "МГУ", Москва, 2013.
21. Васильев Г.Г. "Компьютерные методы анализа текстов." Издательство "Технологии будущего", Санкт-Петербург, 2012.
22. Павлов В.П. "Программные средства для интеллектуального поиска информации." Издательство "Прогресс-Технологии", Москва, 2011.