Министерство образования Республики Беларусь
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.912

Бредихин
Юрий Алексеевич

Автоматизация поиска релевантных текстовых документов в сети интернет

АВТОРЕФЕРАТ
на соискателя степени магистра информационных наук

по специальности 1-40 81 01 “Информатика и технологии разработки программного обеспечения”

Научный руководитель
Калугина Марина Алексеевна
канд. физ.-мат. наук, доцент

Минск, 2018

ВВЕДЕНИЕ

На сегодняшний день объемы хранимой в сети электронной информации растут в геометрической прогрессии. Этому способствует использование мощных СУБД и активное разрастание глобальных сетей. Одна из важных задач, стоящих перед ИПС(информационно-поисковые системы) – это удовлетворение информационных потребностей пользователя: понять, что он планирует получить от выдачи, вводя указанный запрос. Понимание системой намерений пользователя даст ИПС возможность подобрать максимально корректную для этого случая выдачу, тем самым удовлетворив информационную потребность пользователя. В этих условиях к алгоритмам поиска информации, а также к взаимной интеграции программ с различными информационными источниками (в частности, ресурсами сети Интернет) предъявляются все более высокие требования. Именно от производительности поисковых систем – зависит степень полезности многочисленных разрозненных данных, находящихся в сети. Сами по себе – данные не несут пользы до тех пор, пока не станут отвечать контексту конкретного запроса. Однако интенсивный рост количества ресурсов, доступных пользователям в сети Интернет обусловил ряд проблем, перед которыми столкнулись действующие поисковые системы.

Одной из таких проблем является изобилие разнородной информации, значительно затрудняющее поиск. Задача поиска упрощается, если речь идет об редком узкоспециализированном документе со специфической терминологией. Однако, зачастую, на вход поискового механизма подается ключевое слово, характеризующее общее понятие. В этом случае эффективность ранжирования найденных ресурсов заметно снижается. Высокая релевантность в сочетании с низким, а точнее – неконтролируемым уровнем пертитентности выборки – приводит к результатам, допускающим относительно общую типизацию. Поэтому поисковые системы, которые руководствуются пользовательскими ключевыми словами, работают недостаточно эффективно.

Таким образом, можно выделить проблему, возникающую на этапе формирования поискового запроса, которая звучит, как необходимость ручного

анализа текстового документа и правильное составление поискового запроса. Это требует значительных временных затрат на прочтение, подробный анализ и формулировку поискового запроса.

Так же ИПС все еще выявляют потребность в адаптации к информационным запросам, по ключевым словам, сформулированным конкретными пользователями, а также необходимость проведения углубленного поиска по интересующим их тематикам. Из этого следует, что для решения данной проблемы необходимо формировать запрос на основе целого текстового документа.

Традиционные подходы, когда запрос формируется на основании целого документа, опираются, главным образом, на локальные базы текстовых документов, что не позволяет достичь приемлемого результата для пользователя ввиду ограниченности информации. Главная проблема локальных поисковых систем заключается в вопросе расширения индексной базы и ее регулярного обновления для того, чтобы она была согласована с реальной действительностью.

Все это позволяет сделать вывод, что задача разработки алгоритмов и программных средств автоматического поиска текстовых документов в глобальных компьютерных сетях, когда в качестве входного поискового запроса выступает текстовый документ, является актуальной задачей с научной и практической точки зрения.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объект диссертационного исследования – системы автоматического поиска текстовых документов, релевантных данному.

Целью работы является разработка и программная реализация алгоритма автоматического поиска в сети Интернет текстовых документов, релевантных данному.

В соответствии с поставленной целью основными задачами диссертации являются следующие задачи:

1. Исследование требований поиска существующих информационно-поисковых систем в сети интернет.

2. Разработка алгоритмов, предназначенных для определения языка исходного текстового документа.

3. Разработка алгоритмов, предназначенных для выделения ключевых слов.

4. Разработка архитектуры и программных средств автоматического поиска в сети Интернет текстовых документов, релевантных данному.

5. Проведение экспериментальных исследований разработанных моделей, алгоритмов и программ поиска.

Научная новизна. К основным результатам работы, отличающимся научной новизной относятся:

1. Методы, алгоритмы и экспериментальное программное обеспечение процесса определения языка исходного текстового документа.

2. Модели, методы, алгоритмы и экспериментальное программное обеспечение процесса автоматического выделения ключевых слов.

3. Комплексное решение задачи автоматического поиска в сети интернет текстовых документов релевантных данному.

Результатом работы является веб-приложение для Windows, позволяющее автоматически по заданным текстовым документам, находить в сети интернет релевантные текстовые документы.

Область применения результатов: классификация и анализ текстов, информационный поиск.

Личный вклад соискателя. Все изложенные в диссертации результаты исследования получены соискателем лично с учетом замечаний и рекомендаций научного руководителя.

Публикации. Основные положение работы и результаты диссертации изложены в двух опубликованных работах общим объемом 6 с.

Структура и объем работы. Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введения, трёх глав и заключения, библиографического списка и приложений. Общий объем диссертации − 79 страниц. Работа содержит 12 рисунков, библиографический список включает 30 наименований, 1 приложение.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы автоматизированного поиска в сети интернет текстовых документов, релевантных данному. Дано обоснование актуальности темы диссертационной работы и определены основные направления исследований.

В **общей характеристике работы** сформулированы ее цель и задачи, даны сведения об объекте исследования и обоснован его выбор, представлены положения, выносимые на защиту, приведены сведения о личном вкладе соискателя, апробации результатов диссертации и их опубликованности, а также, структура и объем диссертации.

В первой главе рассматриваются средства лингвистического анализа содержимого текстового документа для поиска релевантных ему. Сделан вывод, о том, что большинство современных информационно-поисковых систем используют векторную модель представления данных. Были проанализированы аспекты полноты информационного поиска и существующие подходы к решению задач автоматического поиска документов, которые используют технологии, ориентированные на локальные репозитории. Было определено, что отсутствует возможность поиска в сети Интернет, когда в качестве запроса выступает текстовый документ. На основании проведенного анализа были сформулированы основные задачи, требующие решения.

Во второй главе был проведен анализ информационно-поисковой системы Google, описан алгоритм автоматизированного определения языка текстового документа – разработан базовый алгоритм решения задачи распознавания ключевых слов, основанный на лексико-грамматическом подходе. Предложены улучшения базового алгоритма распознавания ключевых слов – за счет приведения всех слов к каноническим частям речи. Был рассмотрен метод OkapiBM25 для упорядочивания документов по их релевантности. В результате была сформулирована принципиальная схема решения поставленной задачи, решение которой приведено в главе 3.

В третьей главе описаны технологии, используемые для реализации разработанного алгоритма и методика применения, основанного на нем приложения. Приведены результаты анализа результатов работы разработанного поискового программного комплекса.

ЗАКЛЮЧЕНИЕ

1. Проведена систематизация современных теоретических и практических подходов в области автоматизированного поиска текстовых документов, на основе которой выявлены ключевые факторы необходимые для решения поставленной задачи. Выявлены недостатки существующих моделей, алгоритмов и промышленных систем поиска текстовых документов. Произведена оценка эффективности современных методов поиска в ИПС.

2. Разработаны двухэтапная модель определения языка текстового документа и лексико-грамматическая модель выделения ключевых слов, отличающаяся от уже существующих дополнительными этапами лемматизации и тегирования. Произведен более точный анализ каждого полученного документа.

3. Разработана и программно реализована автоматизированная система поиска текстовых документов, релевантных данному. Описаны технологии, используемые при реализации алгоритмов.

4. Приведены результаты экспериментальных исследований. Определены факторы времени работы приложения. Полученные данные, позволяют оценить эффективность и адекватность предложенных и разработанных методов, направленных на поиск релевантных документов в сети Интернет.

5. Проведена апробация проведенной в рамках диссертационного исследования работы, показавшая эффективность используемого метода поиска в разработанной автоматизированной системе.

СПИСОК ПУБЛИКАЦИЙ

Материалы научных конференций

1-А.Бредихин, Ю. А. Автоматическое распознавание ключевых слов в текстовых документах / Ю. А. Бредихин // Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов (Минск, 2 – 6 мая 2017 г.). – Минск: БГУИР, 2017. – С. 162 – 163. URL: https://libeldoc.bsuir.by/handle/123456789/13084 (дата обращения: 20.12.2017)

Статьи

2-А.Бредихин, Ю. А. Автоматическая идентификация языка документа для последующего cross-language анализа // Студенческий: электрон. научн. журн. 2017 № 19 (19). – С. 24 – 29 URL: https://sibac.info/journal/student/19/89477 (дата обращения: 20.12.2017)