Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК [004:005.922]+004.032.2

Семёнова
Дарья Александровна

МЕТОДЫ ПОИСКА ТЕКСТОВЫХ ДОКУМЕНТОВ В СЛАБОСТРУКТУРИРОВАННЫХ ИНФОРМАЦИОННЫХ МАССИВАХ

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1 - 40 80 02 Системный анализ, управление и обработка информации (по отраслям)

Научный руководитель

Кукин Дмитрий Петрович,
кандидат технических наук,
доцент

Минск 2020

ВВЕДЕНИЕ

Электронный документооборот – один из главных аспектов качественного и бесперебойного функционирования любой компании, организации или государственной структуры. Работа с накопленными данными, необходимость быстрой обработки запросов и повышение производительности труда сегодня являются важной стороной работы структуры любого уровня. На тему доступности информации руководителями учреждений проводится множество исследований. Работодатели пытаются оценить в денежном выражении время, потраченное их сотрудниками на поиск нужной информации. Помимо финансовой стороны вопроса рассматриваются качество и эффективность использования уже накопленной информации. Зачастую принимается решение о повторном создании документов, а не о перспективах их обнаружения в имеющейся базе данных.

Такое положение более характерно для малых структур, не обладающих полноценными корпоративными базами данным, позволяющими осуществлять быстрый запрос и поиск необходимой информации. Как правило, данные в них хранятся в слабоструктурированных информационных массивах, не имеющих развитой системы поиска, что не позволяет осуществить своевременный доступ к информации.

При постановке задачи поиска в слабоструктурированных информационных массивах необходимо применять методы, обладающие небольшой сложностью, высоким быстродействием и модифицированные под конкретные условия применения. Тогда полученная система будет характеризоваться минимальным объемом памяти, занимаемый на сервере, гибкостью алгоритмов и хорошим качеством поиска.

Таким образом, практическая значимость работы заключается в создании ИПС для слабоструктурированных информационных массивов, позволяющей повысить быстродействие поиска и учесть основные атрибуты документов за счет применения предметно-ориентированных словарей; изменения границ разрешающей способности при статистическом анализе для увеличения/уменьшения количества термов в анализе; ранжирования документов в выдаче при запросе пользователя; поиска документов как по одному, так и по нескольким атрибутам одновременно.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель исследования

Целью диссертационной работы является повышение быстродействия и произведение отбора параметров поиска в слабоструктурированных информационных массивах на основе разработки аналитической и процедурных моделей информационно-поисковых систем.

Задачи исследования

1 Провести анализ предметной области.

2 Разработать аналитическую модель ИПС в слабоструктурированных информационных массивах.

3 Разработать процедурную модель поиска в слабоструктурированных информационных массивах.

4 Разработать процедурную модель сравнения битовых последовательностей индексов термов.

5 Разработать блочно-модульную структуру ИПС.

Личный вклад соискателя

Соискателем выполнены все изложенные в работе разработки и исследования. Постановка задач и обсуждение результатов проводились совместно с научным руководителем, доцентом кафедры систем управления Белорусского государственного университета информатики и радиоэлектроники. Обработка, интерпретация данных, а также выводы сделаны автором самостоятельно.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались на следующих научных конференциях:

– 55-я юбилейная научная конференция аспирантов, магистрантов и студентов" учреждения образования "Белорусский государственный университет информатики и радиоэлектроники"

СОДЕРЖАНИЕ РАБОТЫ

Сегодня актуальность задачи информационного поиска не вызывает сомнения, кроме того, она тем выше, чем больше объем накопленных неструктурированных данных. Очевидно, что минимизация временных и вычислительных затрат на поиск нужной информации является основным вектором развития данной отрасли знаний. Современные информационные технологии в области анализа данных направлены на обработку массивов больших данных. Основной предпосылкой для развития этих технологий является необходимость обработки большого объема накопленных данных и извлечения из них информации, полезной для выбора стратегии развития и разработки бизнес–моделей.

В первом разделе диссертационного исследования был произведен анализ существующей существующих решений на предмет обеспеченности информационно-поисковыми системами слабоструктурированных информационных массивов. Также были проанализированы существующие подходы к организации информационного поиска в локальных и глобальных сетях. Рассмотрены методы и модели по основным этапам информационного поиска, выявлены основные проблемы и направления исследования при организации информационного поиска в слабоструктурированных информационных массивах.

Во втором разделе был сделан акцент на учет особенностей поиска текстовых документов в слабоструктурированных информационных массивах. Разработана аналитическая модель поиска документов, которая сформирована в виде кортежа, где сам документ представляется вектором из множества представлений документа, который включает в себя пять атрибутов. Разработана процедурная модель сравнения битовых последовательностей индексов термов. Описаны начальные этапы процедурной модели (лингвистический, статистический и кластеризация термов) обработки текстовых документов в слабоструктурированных информационных массивах.

В третьем разделе предметом исследования стала та часть организации поиска в слабоструктурированных информационных массивах, которая непосредственно ориентирована на описание документов коллекции в векторном пространстве, организацию поиска по выделенным атрибутам документа, а также оценку качества разработанной ИПС стандартными для систем поиска в целом параметрами.

В четвертом разделе приведено подробное описание содержания и процесса функционирования разработанной информационно-поисковой системы. Приведено описание состава приложений, реализующего основные этапы поиска – обработки текстовой коллекции документов и непосредственного поиска документов. Показан процесс взаимодействия пользователя и информационно- поисковой системы.

ЗАКЛЮЧЕНИЕ

В процессе написания диссертационной работы были изучены основные алгоритмы поиска в слабоструктурированных информационных системах, построена аналитическая модель поиска текстовых документов, учитывающая атрибуты документов. Построена процедурная модель поиска, включающая дополнительные этапы лингвостатистического анализа и кластеризацию термов на основе уменьшенного до одного байта формата сигнатуры.

В процессе работы были получены основные результаты:

– применение разработанной ИПС позволяет поэтапно сократить количество термов, участвующих в анализе на 31-39 %.

– применение ИПС позволяет сократить время поиска на 18-29 % в случае безошибочного запроса и на 12-20 % в случае запроса с ошибкой в атрибутах «заголовок» или «содержание».

– ИПС позволяет осуществить поиск документов как по одному, так и по нескольким атрибутам одновременно.

Разработанные модели целесообразно применять в компаниях и структурах, где нет организованного хранения документов. В качестве направлений дальнейшего исследования возможно рассмотреть поиск при наличии двух ошибок в запросе пользователя; изучить влияние средней длины терма в предметно-ориентированных коллекциях документов на характеристики поиска.

Таким образом, была получена качественная поисковая система, удовлетворяющая потребности пользователей и соответствующая заявленным требованиям.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Семёнова Д.А. Аналитические и процедурные модели для информационной системы распознавания графических объектов / Д.А. Семёнова, Д.П. Кукин// Сборник докладов 56-й научно-технической конференции аспирантов, магистрантов и студентов БГУИР, Минск, 21-24 апреля 2020 г./БГУИР. – Минск, 2020. – С.