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

УДК 004.55

Ковалёв
Максим Александрович

Алгоритмы поиска видеоинформации по заданному запросу на основе анализа веб-страниц

АВТОРЕФЕРАТ
на соискание степени магистра информатики и вычислительной техники

по специальности 1-40 81-03 Искусственный интеллект

Научный руководитель
Колб Дмитрий Григорьевич
кандидат технических наук

Минск 2015

ВВЕДЕНИЕ

Специфические особенности Веб обуславливают необходимость новых исследований в области организации доступа к информации. Многие новые методы поиска основаны на использовании не только информации о текстовом (тематическом) содержимом документов, но также пытаются использовать другую доступную информацию.

Одним из полезнейших источников такой информации является структура графа Веб, построенного на основе существующих гиперсвязей между страницами. Информация о структуре гиперссылок, является основой алгоритма ранжирования страниц Page Rank.

Другим дополнительным источником информации является HTML-разметка документов. Эта информация может быть использована для вычисления значимости ключевых слов в зависимости от контекста их использования.

Как известно, в настоящее время Интернет представляет собой один из самых больших и постоянно развивающихся источников разнообразных сведений.

Конечно же, сейчас во всемирной паутине существует огромное количество разнообразных поисковых систем, которые в той или иной степени решают задачу поиска. Самыми известными и популярными в России являются поисковые системы Google и Яндекс. Но они помогают человеку в поиске лишь частично, потому что предоставляют ему не ту информацию, которую он желал найти, а ту, которая соответствует поисковому запросу.

В некоторых случаях необходимая информация может располагаться не на первых страницах, и пользователь, просто не увидев её, может завершить поиск. По этой причине для улучшения процесса поиска необходимо производить проверку результатов на соответствие цели поиска.

В данной работе рассматриваются инструменты, необходимые для выполнения эффективного поиска для пользователя, основанные на улучшении результатов выдачи поисковой системы.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации

В настоящее время объёмы данных в сети Интернет стремительно возрастают с каждым днём. Одна и та же информация приобретает различные представления. Этим обуславливается невысокое качество поиска необходимых видео-документов, а соответственно и значительные временные затраты на дополнительный поиск в результатах выдачи поисковых систем.

Исследование предполагает изучение существующих алгоритмов интеллектуального анализа данных, технологий получения результатов поиска у современных поисковых систем, а так же способов улучшения результатов поиска на основе алгоритмов интеллектуального анализа данных. Конечная цель – получение системы с улучшенными качественными показателями поиска на основе ранее изученных методик. Обязательным этапом исследования является моделирование ситуации, демонстрирующей данные алгоритмы и методики.

Цель и задачи исследования

Цель магистерской диссертации – улучшение результатов выдачи поисковой системы за счёт разработки методики поиска видеоинформации

Работа над диссертацией включает в себя следующий список задач:

- Анализ существующих алгоритмов интеллектуального анализа данных
- Анализ технологий получения результатов поиска у современных поисковых систем
- Анализ существующих систем по улучшению результатов поиска на основе алгоритмов интеллектуального анализа данных
- Разработка методики улучшения качества поиска одной из крупных поисковых систем
- Реализация алгоритма поиска видеоинформации по популярной музыкальной тематике для апробации методики поиска видеоинформации
- Оценка полученных результатов

Объект и предмет исследования

Объект исследования – алгоритмы интеллектуального анализа данных

Предмет исследования – методы улучшения качества поиска крупных поисковых систем

Гипотеза

В ходе работы над диссертацией было выдвинуто предположение, что рейтинг страницы, содержащей видеоинформацию, в социальных сетях является основополагающим фактором при ранжировании результатов. Что было подтверждено результатами исследования.

Методология и методы проведенного исследования

Большинство работ в области поиска видеоинформации основаны на поиске сходства между примитивами и ключевыми кадрами видеофайлов, что неприменимо в условиях онлайн-поиска в сети интернет. В нашем случае, когда речь идёт о поиске в сети интернет, целесообразно использовать технологии ИАД для решения поставленной задачи.

На данный момент существуют некоторые методы решения задач подобного класса: расширение запросов пользователей, классификации текстовой информации, ранжирование выборки данных.

Рассмотренные выше методы в исходном виде применимы только для текстовой информации, где основное ранжирование результатов происходит по весу текстов на основе ключевых слов запроса.

Стоит отметить, что поиск в интернете, не может быть корректно выполнен, будучи основан на анализе одного лишь текста документа. Ведь факторы, отличные от текстовых, играют не меньшую, а порой и большую роль, чем текст самой страницы. Положение на сайте, посещаемость, авторитетность источника, частота обновления, цитируемость страницы и ее авторов – все эти значения должны учитываться при осуществлении поиска. А при осуществлении поиска видеоинформации нет смысла рассматривать страницы, не имеющие плееров или контейнеров для воспроизведения видео в своём коде.

Научная новизна и значимость полученных результатов

Научная новизна и значимость результатов заключается в том, что предметная область рассматривается сразу с разных ракурсов обработки информации, от низкоуровневого к высокоуровневому, а совместное

использование разноуровневых методов повышает результативность работы алгоритмов каждого уровня в отдельности.

Усовершенствование алгоритмов ранжирования позволяет обратить внимание на новые значимые составляющие пертинентности результатов поиска для конечного пользователя, что является отправной точкой для новых исследований в этой области.

Практическая (экономическая, социальная) значимость полученных результатов

Разработанная система может быть использована в качестве новой поисковой системы по видеоматериалам в сети интернет, а так же на её базе можно разработать систему, предоставляющую API для приложений, нуждающихся в выполнении поиска видеоинформации.

Результаты диссертации позволяют оценить значимость критериев поиска в ранее недостаточно изученном контексте поиска видеоинформации в сети Интернет.

В качестве направлений дальнейшего исследования можно выделить следующие:

- совершенствование методов полнотекстового поиска
- интеграция различных подходов при оценке релевантности страницы
- использование нескольких поисковых систем для обеспечения полноты поиска
- реализация API для возможности встраивания системы в другие проекты

Основные положения диссертации, выносимые на защиту

1. Предложен алгоритм ранжирования, улучшающий результаты поиска.
2. Разработана методика улучшения качества поиска на основе алгоритмов интеллектуального анализа данных.
3. Актуализация информации за счёт циклического автоматического извлечения данных из результатов поисковой системы, предложенных к выдаче.
4. Разработан программный продукт, позволяющий производить поиск видеоинформации по заданному запросу.

Личный вклад соискателя

Была разработана формула ранжирования результатов поиска, а так же произведена комбинация различных подходов структурного анализа, что

позволило улучшить результаты выдачи поисковой системы Яндекс. Был разработан программный продукт, исходя из поставленных задач.

Апробация результатов диссертации

Доклад по теме работы был представлен на 50-й научной конференции БГУИР (2014 г.) в секции «Интеллектуальные информационные технологии».

Структура и объем диссертации

Общий объем магистерской диссертации составляет 64 страницы, включая 10 иллюстраций, 2 таблицы и библиографический список из 32 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В главе 1 были рассмотрены основные понятия интеллектуального анализа данных, даны определения основным тезисам, проведен обзор существующих проблем. Исследованы в полной мере технологии, алгоритмы и методы проведения анализа.

Глава 2 посвящена обоснованию выбора методов улучшения качества поиска и последующей классификации полученных результатов, рассмотрены метод расширения запросов пользователей, метод классификации текстовой информации, метод ранжирования выборки данных. Проведен анализ средств для решения задачи полнотекстового поиска.

В главе 3 произведён выбор базовой поисковой системы, способ улучшения результатов поиска, а так же рассмотрена архитектура и возможности реализованной программной системы использующий методы, описанные в данной работе. Разработана универсальная формула подсчёта рейтинга каждой отдельно-взятой страницы.

ЗАКЛЮЧЕНИЕ

Поиск информации в Интернете - это еще не до конца изученная проблема. В настоящее время существует масса ее решений. Подход, описанный в данной работе, обладает следующими достоинствами:

- Возможность применения для любой предметной области
- Улучшение пертинентности поиска в разработанной системе
- Актуальность и полнота результатов поиска

В рамках данной работы были получены следующие результаты:

- Проведен анализ информационно-поисковых систем
- Рассмотрены теоритические и практические методы решения задачи улучшения качества поиска
- Разработан процесс взаимодействия с API поисковой системы
- Разработана модель улучшения качества поиска
- Создан алгоритм поиска видеоинформации в глобальной сети на основании полученных результатов поиска
- Реализована разработанная модель системы для апробации методики
- Произведена оценка полученных результатов

По результатам данной работы было получено, что все документы, найденные по запросам, релевантны теме соответствующих запросов и представляют собой качественную выборку для пользователей, использующих данную систему.

К недостаткам системы следует отнести её низкую скорость работы, которая в большей части обусловлена скоростью получения множества ответов от различных ресурсов, а так же зависимость от сторонних сервисов по доставлению данных – в случае, если сервисы исчезнут или закроют публичный API, придётся искать новые способы по извлечению данных.