Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра математического и компьютерного моделирования

Проектирование и разработка информационно-поисковой машины

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы

направления 09.03.03–Прикладная информатика

механико-математического факультета

Сарсеновой Айжан Зинетуллаевны

Научный руководитель
доцент, к.ф.–м.н., доцент С.П. Шевырев

Зав. кафедрой
зав. каф., д.ф.–м.н. Ю. А. Блинков

Саратов 2017

СОДЕРЖАНИЕ

Стр.

ВВЕДЕНИЕ

Данная работа посвящена одной из самых важных областей в сфере науки – информационному поиску и средствам поиска. Информационный поиск, появившийся изначально как отдельная дисциплина библиотечного дела, которая использует средства информатики, некоторое время оставался скромной научной областью, в которой работало небольшое количество ученых. Сейчас информационный поиск – это междисциплинарная наука, стоящая на пересечении информатики, лингвистики, когнитивной психологии, информационного дизайна, семиотики и библиотечного дела. Появление Интернета дало значительный скачок в развитии этой науки. В настоящее время информационный поиск – это миллионы пользователей по всему миру, большие массивы данных, мощные вычислительные системы, сложные алгоритмы. И для того, чтобы осуществлять поиск информации в Глобальной сети, были созданы информационно-поисковые системы.

Информационно-поисковые системы – это основное средство поиска информации. Как правило, принцип работы поисковых систем основан на использовании запросов, состоящих из ключевых слов, передающихся как аргументы поиска, то есть эти слова являются, непосредственно, формулировкой информационной необходимости человека. Сам поиск автоматизирован: он осуществляется с помощью поискового робота и индексатора. И в связи с огромными мощностями поисковых систем и эффективностью алгоритмов выдача результатов выполняется настолько быстро, что обычный пользователь не может за это время достать книгу.

Актуальность выбранной темы заключается в том, что в настоящее время без использования средств и услуг информационно-поисковых машин не обходится ни один пользователь Интернета. Информационно-поисковые системы являются обязательной частью глобальной сети. Сейчас поисковые системы – это не просто огромный и сложнейший механизм, который является инструментом для нахождения любой необходимой информации, но также представляют довольно привлекательную сферу для бизнеса.

Целью данной работы является разработка информационно-поисковой машины, которая может осуществлять поиск по веб-страницам и сортировку

результатов поиска по релевантности найденных документов запросу. **Объектом исследования** является процесс поиска, выполняемого пользователями, необходимой информации в глобальной сети. **Предметом исследования** являются алгоритмы и методы полнотекстового поиска, с помощью которых осуществляется работа информационно-поисковых машин.

Задачами этой работы, которые будут рассмотрены и решены в ходе исследования, станут:

- ознакомление с понятием, задачами и видами информационного поиска;
- описание принципов и особенностей работы поисковой системы;
- знакомство с историей развития поисковых систем;
- изучение архитектуры поисковых машин, их состав;
- ознакомление с типами веб-сервисов, которые предназначены для текстового или графического поиска в Интернете;
- анализ работы поисковой машины.

Характеристика материалов исследования. Для разработки информационно-поисковой машины были выбраны следующие технологии:

- высокоуровневый язык программирования общего назначения Python;
- встраиваемая реляционная база данных SQLite.

Высокоуровневый язык программирования общего назначения Python был выбран для реализации поисковой машины, потому что этот язык является динамически типизированным, и написанный на нем код является короче, чем на других популярных языках. В стандартный дистрибутив Python уже входит много библиотек, в том числе для вычисления математических функций, разбора HTML-страниц и загрузки веб-страниц. Также Python дает возможность тестировать функции, не составляя отдельную тестовую программу. Программы можно запускать прямо из командной строки, и в нем есть интерактивный режим, в котором разршается вызывать функции, создавать объекты и тестировать пакеты. Python поддерживает объектно-ориентированный, процедурный и функциональный стили программирования [1].

Для хранения информации о веб-документах, которые были проиндексированы поисковой машиной, была использована компактная встраиваемая реляционная база данных SQLite. Так как SQLite является встраиваемой,

то значит, что она не использует клиент-серверную архитектуру, то есть не является отдельно работающим процессом, с которым взаимодействует программа, а предоставляет библиотеку, с которой происходит коммуникация с исходной программой, и SQLite становится составной частью написанной поисковой машины. Использование SQLite значительно уменьшает временные ресурсы, время отклика и упрощает программу. Эти качества очень существенны, особенно при работе с огромным массивом данных, с которым приходится работать информационно-поисковой машине. SQLite хранит всю базу данных (включая определения, таблицы, индексы и данные) в единственном стандартном файле на том компьютере, на котором исполняется программа [2].

Структура выпускной квалификационной работы. Бакалаврская работа содержит обозначения и сокращения, введение, 3 раздела: «Информационный поиск», «Информационно-поисковые системы», «Поисковая машина», заключение, список использованных источников из 33 наименований и одно приложение «Программный код поисковой машины». Общий объем работы – 54 страницы.

1 Основное содержание работы

В введении содержится краткое описание информационного поиска (ИП) и информационно-поисковых систем, описываются актуальность и цель работы, объект, предмет и задачи исследования.

Первый раздел «Информационный поиск» содержит характеристику ИП, который состоит из следующих частей:

- определения понятия информационного поиска, описания процесса поиска, формирования запроса и характеристики объекта запроса;
- описания главной задачи информационного поиска; рассмотрен также расширенный список задач;
- рассмотрения видов поиска информации, которые делятся на адресный (поиск документов, которые которые соответствуют «внешним» признакам в запросе [3]), фактографический (поиск фактов, удовлетворяющих запросу, где факты – это конкретные сведения о каких-либо информационных объектах [4]), документальный (процесс нахождения и выдачи документов из хранилища информационно-поисковой системы, которые удовлетворяют запросу пользователя [5]) и семантический (поиск документов по их содержанию [6]).

Второй раздел был посвящен следующим вопросам, которые связаны с поисковыми системами, то есть что они из себя представляют, как устроены, каким образом формируется ответ на поисковый запрос, как происходит ранжирование документов при поиске. Для начала дается описание развития поисковых систем, после содержание второго раздела выглядит следующим образом:

- описывается архитектура поисковой системы, которая состоит обычно состоит из трех компонентов [7]: поискового робота (программа, с помощью которой автоматически перебираются страницы в Интернете для занесения информации в базу данных поисковый системы), индексирования (процесс занесения информации о веб-странице в базу данных поисковой системы), поисковика (графический интерфейс для работы пользователя, обычно состоит из поля, в которое следует ввести свой поисковый запрос);

- рассматриваются типы поисковых систем, которые в соответствии с описанием на электронном ресурсе [8] делятся на четыре типа: системы, использующие поискового робота (системы, которые состоят из трех частей: поискового робота, индекс и программное обеспечение поисковой системы), системы, управляемые человеком или, по-другому, каталоги ресурсов (каталог ресурсов ищет результаты, исходя из описания сайта, которое предоставляется веб-мастером), гибридные системы (системы, сочетающие в себе функции систем, использующих поискового робота, и систем, которые управляются человеком) и мета-системы (системы, которые объединяют и ранжируют результаты сразу нескольких поисковиков);
- приводятся характеристики информационно-поисковых систем, среди которых основными считаются: полнота (отношение количества найденных релевантных документов к общему объему документов, хранящихся в базе данных поисковой системы) и релевантность (соответствие поисковому запросу информации, выдаваемой поисковой системой) [9], но на практике чаще используется понятие пертинентности – это соотношение объема всех документов, являющихся полезными для пользователя, к общему объему документов, выданных в результате поиска [10];
- описывается лингвистическое обеспечение, которое играет важную роль в развитии информационно-поисковых систем (средства лингвистики реализуют интерфейс между естественным языком и формальными поисковыми механизмами ИПС), также приводится описание стоп-слов, которые не играют роли в процессе выявления документов, удовлетворяющих запросу пользователя [11], построения морфемного анализа и тезауроса;
- представляются особенности и принципы работы поисковых систем, из которых следует, что сама поисковая процедура имеет определенную этапность: от определения информационной потребности и области поиска до анализа результатов и выбора пертинентных объектов, где информационные потребности пользователя могут относиться к разным областям, то есть могут иметь широкую направленность или, наоборот, узкоспециализированными [12], причем сформировать эффектив-

ный запрос к информационно-поисковой системе – это основная задача второго этапа, а третий этап зависит от пертинентности найденного решения;

- приводится статистика и рейтинг популярных сетевых информационно-поисковых служб в мире и в России.

В третьем разделе описывается создание поисковой машины. Поисковая машина – это комплекс программ, который предназначен для поиска информации, являющийся частью поисковой системы, заключается в сборе данных (или документов). Обычно для сбора данных применяется краулер (поисковый паук, веб-паук), который начинает свою работу с небольшого количества документов, а потом переходит по имеющимся в них ссылкам, либо поиск начинается с фиксированного набора документов, которые могут храниться в корпоративной сети интранет[1] [1]. В работе используется первый метод. После этого все собранные документы необходимо проиндексировать. Для этого строится таблица, которая состоит из списка документов и количества вхождений различных слов. В зависимости от конкретного приложения сами документы могут и не храниться в базе данных. В индексе хранит ся лишь ссылка (URL[2] или путь в файловой системе) на их местонахождение.

Последний шаг поиска заключается в возврате ранжированного списка документов, который является результатом ответа на запрос пользователя. Если есть индекс, то найти документ, содержащий заданные слова, довольно несложно, проблема заключается в том, как отсортировать полученный результат. Можно использовать огромное количество метрик или придумать свою, и в принципе недостатка в способах настройки изменения порядка документов этих метрик нет, но стоит лишь ознакомиться с ними, как возникает желание, чтобы большие поисковики предоставляли свои средства для более точного результата (например, слова из запроса должны находится в документах рядом). В подразделе «Алгоритм PageRank» рассмотрен алгоритм ранжирования страниц PageRank, в котором учитываются ссылки на оцениваемую страницу с других страниц. И последний подраздел посвящен

[1]Интранет – это внутренняя частная сеть организации. Как правило, интранет – это Интернет в миниатюре, который построен на использовании протокола IP для обмена и использования некоторой части информации внутри этой организации.

[2]URL – единообразный указатель местонахождения ресурса.

результатам работы поисковой машины, для получения которых была выбрана в качестве примера англоязычная версия Википедия. В ходе анализа результатов стало ясным, что наибольший ранг имеет главная страница сайта, но это неудивительно, так как все страницы википедии ссылаются на «Main Page».

В Приложении А представлен программный код поисковой машины, реализованный на языке Python с применением встроенной базы данных SQLite.

ЗАКЛЮЧЕНИЕ

Сейчас в эпоху информационных технологий трудно представить себе мир без компьютера, смартфонов, планшетов и других устройств, с помощью которых возможен выход в Интернет. За все время существования глобальной сети не раз пытались создать механизмы, которые бы организовывали поиск информации. Многие попытки оказались неудачными, а другие же привели к созданию удобных и полезных средств поиска. Информационно-поисковые системы – это мощный механизм поиска информационных ресурсов в глобальной сети и без него не обходится ни один пользователь Интернета.

Подводя итоги моей выпускной квалификационной работы, хочу отметить, что мною были представлены и решены все установленные во введении задачи и за их счет была достигнута цель работы. В рамках предмета исследования, то есть изучения алгоритмов и методов полнотекстового поиска, с помощью которых осуществляется поиск информации, был рассмотрен объект исследования, представляющий собой процесс поиска пользователями необходимой информации в Интернете. Был разобран алгоритм ранжирования PageRank, придуманный создателями Google, и приведен программный код, описывающий работу веб-робота, написанный на языке программирования Python.

Таким образом, было применено знание теоретической базы на практике и был получен опыт решения задач информационного поиска. Хочется также отметить, что важность разработки в области усовершенствования поисковых машин огромнейшая, количество информации и запросов от пользователей растет каждый день, необходимо успевать обрабатывать, сортировать и выдавать информацию, удовлетворяя постоянно растущие запросы пользователей. Необходимо смотреть в завтрашний день, предугадывать нагрузку на поисковую систему в будущем, применять все новые средства обработки и выдачи ранжированной информации. Современные поисковые системы не стоят на месте, постоянно развиваясь, такой подход позволяет заниматься не только постоянной борьбой и конкуренцией, но и уделять внимания развитию дизайна сайта поисковика, удобству интерфейса, получению коммерческой прибыли от проекта.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Сегеран, Тоби. Программируем коллективный разум. — М.: Символ-Плюс, 2012. — С. 256. — ISBN: 5-93286-119-3.

2. СУБД-движок SQLite. — URL: http://bourabai.ru/dbt/dbms/sqlite.htm (дата обращения: 23.04.2017).

3. Адресный поиск как проявление информационной потребности в конкретном документе. — URL: http://ifreestore.net/3629/24/ (дата обращения: 12.04.2017).

4. Информационный поиск в информатике и библиотековедении. — URL: http://cyberleninka.ru/article/n/informatsionnyy-poisk-v-informatike-i-bibliotekovedenii/ (дата обращения: 12.04.2017).

5. Документальный информационный поиск. — URL: http://cyberleninka.ru/article/n/informatsionnyy-poisk-v-informatike-i-bibliotekovedenii (дата обращения: 12.04.2017).

6. Басипов, А. А., Демич, О. В. Семантический поиск: проблемы и технологии // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. — 2012. — Т. 4, № 1. — С. 104–111.

7. Risvik, K. M., Michelsen, R. Search engines and web dynamics // Computer Networks. — 2002. — Vol. 39, no. 3. — P. 289–302.

8. Поисковая система. — URL: https://ru.wikipedia.org/wiki/\T2A\CYRP\T2A\cyro\T2A\cyri\T2A\cyrs\T2A\cyrk\T2A\cyro\T2A\cyrv\T2A\cyra\T2A\cyrya_\T2A\cyrs\T2A\cyri\T2A\cyrs\T2A\cyrt\T2A\cyre\T2A\cyrm\T2A\cyra (дата обращения: 19.04.2017).

9. Гусев, В.С. Освоение в Internet: Самоучитель. — М.: Вильямс, 2004. — С. 336. — ISBN: 5-8459-0467-6.

10. Ландэ, Д.В. О чем говорят запросы пользователей к поисковым серверам // Сети и телекоммуникации. — 1999. — Т. 12, № 4. — С. 19–21.

11. Гращенко, Л. А. О модельном стоп-словаре // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. — 2013. — Т. 150, № 1. — С. 40–46.

12. Ландэ, Д.В. Поиск знаний в Internet. Профессиональная работа. — М.: Вильямс, 2005. — С. 272. — ISBN: 5-8459-0764-0.