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

Применение методов искусственного интеллекта в поисковых системах

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы

направление 09.03.03 – Прикладная информатика

механико-математический факультет

Хламовой Ирины Александровны

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

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

Саратов 2017

ВВЕДЕНИЕ

В настоящее время информационные технологии стремительно развиваются. В связи с этим появляются всё больше информации различного рода. Однако разобраться в этом огромном обьеме данных самостоятельно - это достаточно трудоемкий и требущий длительного времени процесс. Поэтому поиск по этим реусурсам выполняется с помощью поисковых систем.

В данной работе рассматривается одна из наиболее важных сфер научной жизни - информационный поиск.

Информационный поиск (англ. information retrieval) — процесс поиска неструктурированной документальной информации, удовлетворяющей информационные потребности пользователей, а также - наука о данном поиске.

Исследования, связанные с поиском информации, начались еще в середине прошлого века. В результате были разработаны сотни поисковых систем. Однако практическая реализация в системе поиска информации нового поколения тесно связана с использованием методов искусственного ителлекта.

Главное требование к любой поисковой системе – быстрое формирование точного и полного ответа, адекватного запросу пользователя. Поэтому изучение построения (программирования) поиска, ее методов, проблем, возникающих при поиске информации, является до сих пор актуальной темой.

Основная цель работы заключается в разработке вспомогательного элемента поисковых систем - алгоритма ранжирования результатов, с помощью которого осуществляемый поиск становится более эффективным.

Для достижения данной цели, необходимо решить следующие задачи:

1. Выявление существенных факторов, влияющих на ранжирование;
2. Оценка меры похожести поискового запроса и документов из коллекции;
3. Разработка и программная реализация алгоритма PageRank;
4. Разработкаи и программная реализация алгоритма TextRank.

Работа делится на несколько разделов, в которых рассматриваются задачи, виды и методы информационного поиска, принципы и особенности поисковых систем, а также детальнее изучается один из компонентов системы –

вспомогательный алгоритм ранжирования PageRank и его приложение - алгоритм TextRank, демонстрируются действия этих алгоритмов на конкретном примере.

Основным результатом работы является разработанный на языке программирования PHP код, содержащий реализацию алгоритма PageRank, с помощью которого можно эффективно ранжировать результаты поиска

Основное содержание работы

Первый теоретический раздел содержит разделение на два подраздела: «Задачи информационного поиска» и «Виды и методы информационного поиска»

В первой части раздела содержатся общие сведенья об информационном поиске, в частности рассматриваются описывающий его алгоритм, и основные задачи.

Поиском информации или информационным поиском(ИП) называется некая последовательность операций во множестве документов, выполняемых с целью отыскания документов, содержащих определенную информацию (с последующей выдачей самих документов или их копий), или с целью выдачи фактических данных, представляющих собой ответы на данные вопросы. Все найденные за много лет средства и приемы поиска информации доступны и эффективны и при поиске информации в Интернете. ИП производится при помощи информационно-поисковых систем (ИПС). ИПС - автоматизированная поисковая система, реализованная на средствах электронной вычислительной техники и предназначенная для нахождения, а также выдачи ее пользователям необходимой информации по заданным критериям. ИПС представляет собой совокупность информационно-поискового языка, программных средств и правил перевода текстов на этот язык, обеспечения их поиска и критериев соответствия.

Процесс информационного поиска на общем уровне описывается следующим алгоритмом [1]:

1. Уточнения информационной потребности и формулировка запроса, выделение в его структуре основных поисковых признаков: ключевых слов и понятий, предметов и аспектов поиска.
2. Идентификация данных: сравнение поисковых признаков с данными в информационном (поисковом) массиве.
3. Отбор: проверка выявленного подмассива документов или данных на соответствие заданным критериям поиска.
4. Структурирование (упорядочение) документов или данных в соответствии с логикой запроса.

Центральной задачей информационного поиска является удовлетворение потребности пользователя в инофрмации, сформулированной в запросе. Информационный запрос - это словесное выражение определенной информационной потребности пользователя системы. Запросы анализируются по своему предметному содержанию и описываются в терминах, отобранных из словаря конкретного информационно-поискового языка.Кроме специального языка запросов, современные поисковые системы позволяют вводить запрос на естественном языке.

Однако перечень задач информационного поиска регулярно расширяется и теперь содержит:

- фильтрация документов;
- классификация документов (отнесении документа к одной из нескольких категорий на основании содержания документа);
- вопросы моделирования;
- кластеризация документов (автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества документов);
- проектирование архитектур поисковых систем и пользовательских интерфейсов;
- извлечение информации, в частности аннотирования и реферирования документов;
- языки запросов и др.

Во второй части раздела собрана информация о видах и методах, используемых в информационном поиске.

Существует несколько классификаций по видам информационного поиска. Например, классификация видов поиска по способу применения:

- *Полнотекстовый поиск* - самый простой вид поиска, при котором поиск информации производится во всем объеме данных - по всему объему текста или по всем полям базы данных. И это его главное преимущество. Здесь не нужно знать, как и где хранится информация, просто осуществляется ее поиск. Существенный недостаток данного поиска - уменьшение скорости поиска при увеличении объема данных. Это делает невозможным применение такого механизма для поиска информации в достаточно большой структуре данных.
- *Поиск с запросом* - под данным типом поиска понимается поиск информации в базе данных. Этот тип поиска наиболее универсальный, поскольку с помощью него можно производить поиск информации в огромных базах данных.
- *Инкрементный поиск.* Идея данного вида заключается в том, что поиск осуществляется после каждого нажатия на клавишу, при котором происходит изменение искомой строки - при обычном поиске сначала вводится строка поиска, а затем при нажатии на клавишу "Enter"или кнопку "Найти"запускается механизм поиска. Инкрементный поиск - быстрый поиск с постепенным уточнением.

Были рассмотрены методы информационного поиска. Различают:

- в зависимости от цели - *адресный* (формально-механический) и *семантический* (тематический);
- от объекта поиска - *документальный* и *фактографический*;

Адресный поиск — это поиск данных о наличии и/или местонахождении, точном адресе хранения документа. Процесс поиска документов по чисто формальным признакам, указанным в запросе: точный адрес документа или адреса хранения документов в хранилище. Такой поиск сведений о наличии и местонахождении конкретного документа часто и называется библиотечным.

Семантический поиск - поиск данных, при котором процесс идентификации информации происходит по её содержанию. [2]

Документальный поиск заключается в нахождении в хранилищах информационно-поисковой системы конкретных первичных документов (биб-

лиотечный вид документального поиска) или в базах данных вторичных документов (библиографический вид документального поиска), которые соответствуют запросу пользователя.

Фактографический поиск - это вид информационного поиска, связанный с процессами нахождения и выдачи конкретных (фактографических) данных. Фактографический поиск — это поиск какого-либо конкретного факта, данных, относящихся к какому-либо предмету, процессу, событию; поиск терминов, законов, дат, адресов, правил правописания и т. д., и т. п. Конечным результатом фактографического поиска является не документ, не список документов, а ответ по существу.

Различают два вида:

- Документально-фактографический, заключается в поиске в документах фрагментов текста, содержащих факты.
- Фактологический (описание фактов), предполагающий создание новых фактографических описаний в процессе поиска путём логической переработки найденной фактографической информации.

Во втором разделе ссодержится основная информация о системах искусственного интеллекта и области применения.

Искусственный интеллект определяется как научная дисциплина, которая занимается моделированием разумного поведения. Центральные задачи искусственного интеллекта состоят в том, что бы сделать вычислительные машины более полезными и чтобы понять принципы, лежащие в основе интеллекта.

Выделяютмя два направления развития искусственного интеллекта:

- решение проблем, связанных с приближением специализированных систем искусственного интеллекта к возможностям человека, и их интеграции, которая реализована природой человека;
- создание искусственного разума, представляющего интеграцию уже созданных систем искусственного интеллекта в единую систему, способную решать проблемы человечества.

В основном выделяются следущие направления искусственного интеллекта, которые решают задачи, что плохо поддаются формализации: доказательство теорем, распознавания изображений, машинный перевод и понимание

человеческой речи, игровые программы, машинная творчество, экспертные системы. Далее кратко рассматривается их сущность.

А так же приводятся примеры самых известных систем исскуственного интеллекта и описывается их деятельность .

Deep Blue — победил чемпиона мира по шахматам. Матч Каспаров против суперЭВМ не принёс удовлетворения ни компьютерщикам, ни шахматистам, и система не была признана Каспаровым. Watson — перспективная разработка IBM, способная воспринимать человеческую речь и производить вероятностный поиск, с применением большого количества алгоритмов. MYCIN — одна из ранних экспертных систем, которая могла диагностировать небольшой набор заболеваний, причем часто так же точно, как и доктора. 20Q — проект, основанный на идеях ИИ, по мотивам классической игры «20 вопросов». Стал очень популярен после появления в Интернете на сайте 20q.net Распознавание речи. Системы такие как ViaVoice способны обслуживать потребителей. Роботы в ежегодном турнире RoboCup соревнуются в упрощённой форме футбола.

Третий теоретический раздел состоит из описания видов поисковых систем, а так же алгоритма их действия.

Основная задача поисковой системы - минимизировать время, затрачиваемое пользователем на поиск релевантной запросу информации. Релевантность - одно из самых субъективных и запутанных понятий в науке информационного поиска. Наиболее часто говорят о релевантности с точки зрения пользователя, и тогда «релевантная запросу информация» и «нужная пользователю информация» - одно и то же.

Все поисковые системы можно разделить на группы по некоторым критериям.

По способу работы поисковые системы делятся на:

- поисковые каталоги (Yahoo!, Lycos, www.list.ru, www.ulitka.ru)
- поисковые указатели (Google, Rambler)

Как правило, системы работают поэтапно:

1. поисковый робот получает контент (содержимое, информационное наполнение сайта);
2. индексатор генерирует доступный для поиска индекс;

3. поисковик обеспечивает функциональность для поиска индексируемых данных.

В четвертом разделе содержится описание алгоритма PageRank с использованием графов и его программная реализация.

PageRank(PR) - это алгоритм оценки сайта поисковой системой Google по количеству и качеству внешних ссылок на данный ресурс. Другими словами - это мера «важности» и «авторитетности» страницы. Понятие PageRank является одним из ключевых моментов в работе поисковой машины. PR является одним из вспомогательных факторов при ранжировании сайтов в результатах поиска. Алгоритм применяется к коллекции документов, связанных между собой гиперссылками. Чаще всего PageRank используется для ранжирования веб-страниц. Однако знания работы PR можно приложить и при работе с любым набором объектов, связанных между собой взаимными ссылками (например к графу). Высокое значение PageRank обеспечивает сайту хорошее положение при выдаче результатов поиска в Google, следовательно, получение высокого PageRank является одной из основных задач поисковой оптимизации.

Алгоритм PageRank рассматривает Интернет как ориентированный граф, в котором страницы - это узлы, а гиперссылки - связи между этими узлами. Он может быть использован для ранжирования узлов любого вида графа (в том числе и неориентированных) по степени важности.

Поскольку нахождение значение вероятности PageRank для отдельного узла в графе будет зависеть от значения PageRank всех узлов, которые соединены с ней, и эти узлы могут подключаться циклически к узлу, чьи ранжирования мы хотим получить, значения PageRank обычно назначаются с помощью итерационного метода.

Поскольку PageRank с использованием графов сводится к вычислению показателя как в базовом алгоритме, то приведем выжимку из кода, в которой содержится объявление вектора документов **a** и матрица смежности *ms*, а также определение с помощью матрицы, на какой документ данный экземпляр ссылается и какие документы ссылаются на него.

```
<?php
$N = 4;
```

```
$a = array(

1 => 'doc1',
2 => 'doc2',
3 => 'doc3',
4 => 'doc4'

);

$link = array(
    1 => array(),
    2 => array(),
    3 => array(),
    4 => array()
);

$ms = array(
        [0, 0, 1, 1],
        [1, 0, 1, 1],
        [0, 1, 0, 1],
        [0, 0, 1, 0]
);


for ($i=1; $i<=$N;; $i++)
{
        $k=$a[$i];
        for ($j=1; $j<=$N; $j++)
        {
                if ($ms[$i][$j] == 1){
                        $link[$k] => array_push('doc'.$j);
                }

        }
}
?>
```

Так же модель графа была применена к обработке естественного языка, приводящей к алгоритму под названием TextRank. TextRank — приложе-

ние алгоритма PageRank к задачам обработки естественного языка. Основная идея заключается в выполнении трёх шагов:

- построение графа на основе исходного текста на естественном языке;
- приближённое вычисление значения PageRank для построенного графа;
- применение полученных весов вершин для извлечения сведений из текста.

В общем виде, величина TextRank — это значение стационарного распределения случайного блуждания для каждой вершины $t \in V$ с учётом весов рёбер [9].

$$TR(t_i) = (1 - d) + d * \sum_{t_j \in In(t_i)} \frac{\omega_{ji}}{\sum_{t_k \in Out(t_j)} \omega_{jk}} * TR(t_j),$$

где d — фактор затухания;

In(t) — множество вершин, входящих в t;

Out(t) — множество вершин, исходящих из t;

w_{ij} — вес ребра (t_i, t_j).

ЗАКЛЮЧЕНИЕ

В ходе работы были исследованы классическая и модифицированная версии алгоритма PageRank, применяемого для ранжирования результатов в поисковых системах, а также его приложение - алгоритм TextRank.

Цели, сформулированные во введении, достигнуты. Получены теоретические знания в следующих областях:

- информационный поиск;
- поисковые системы;
- принципы действия алгоритма ранжирования PageRank и его приложения - TextRank.

На практике более детально изучалось применение PR при решении реальных задач информационного поиска. Программная реализация классической и модифицированной версий алгоритма также была приложена к данной курсовой работе.

Подводя итоги, можно сделать вывод о широких возможностях применения алгоритма ранжирования PageRank, т.к. он позволяет сделать информационный поиск более эффективным, по сравнению с иными поисковыми системами.

Все проведенные в работе исследования позволили изучить процесс информационного поиска в поисковых системах, использующих алгоритм ранжирования PageRank, а также повысить свои навыки решения задач в данной области как теоретическим, так и практическим способами.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Фишкин, А. В. Поиск в интернете. — Альфа-Пресс, 2014. — С. 105.

2. Колисниченко, Д. Н. Поисковые системы и продвижение сайтов в Интернете. — М.: Диалектика, 2007. — С. 272. — ISBN: 978-5-8459-1269-5.

3. Маннинг, К., Рагхаван, П., Шютце, Х. Введение в информационный поиск. — М: Вильямс, 2011. — ISBN: 978-5-8459-1623-5.

4. Вайз, Д.А., Малсид, М. Google. Прорыв в духе времени. — М.: Диалектика, 2007. — С. 368. — ISBN: 978-5-699-22216-2.

5. Романенко, В.Н. Сетевой информационный поиск. — СПб.: Профессия, 2003. — С. 283.

6. Как работает PageRank. — 2008. — URL: `https://sites.google.com/site/pagerankclub/matematiceskij-algoritm-pagerank`.

7. Шкондин, А. PageRank: Больше ссылок хороших и важных. — 2001. — URL: `http://www.developing.ru/seo/pagerank.html`.

8. Brin, S., Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. — 2003. — URL: `http://infolab.stanford.edu/~backrub/google.html`.

9. The PageRank algorithm. — 2003. — URL: `http://dpk.io/pagerank`.

10. Люгер, Дж. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. — М.: Вильямс, 2005. — С. 864. — ISBN: 5-8459-0437-4.

11. Нильсон, Н. Искусственный интеллект. — М.:Мир, 1973. — Р. 273.