Информационно-поисковый язык и индексирование

Студент: Латиф Гулиев
Факультет: ITT
Группа: 640R
Преподаватель: Мехрибан Фаттахова

Тема: Информационно-поисковые системы

План

1. Понятие информационно-поисковой системы.
2. Виды поисковых средств в Интернете.
3. Характеристика поисковой системы Интернета. Информационно-поисковый язык.

1. Понятие информационно-поисковой системы.

Традиционными ***способами фильтрации и отбора информации*** человеком являются:

- ✔ поиск «сверху» (по оглавлению);
- ✔ поиск «снизу» (с помощью различных указателей);
- ✔ поиск с помощью гипертекстовых связей (перекрестных ссылок);
- ✔ полнотекстовый поиск путем просмотра всего текста.

Организация поиска предполагает следующие составляющие и этапы:

1) множество документов (текстов или их фрагментов), по которым следует производить поиск;
2) коммуникативная потребность в информации, выражающаяся в информационном запросе пользователя;
3) удовлетворение коммуникативной потребности, состоящее в выборе той части текстов исходного массива, которая соответствует информационному запросу.

Информационно-поисковая система (ИПС) - упорядоченная совокупность документов и информационных технологий, предназначенных для хранения и поиска информации, представленной в виде текстов или их частей (фактов)

Для экономии усилий человека с 1950-х годов осуществляются попытки создания автоматизированных ИПС. При этом в первых ИПС анализ и описание содержания документов (индексирование) выполнялись вручную, а поиски по этим документам проводились автоматически.

2. Виды поисковых средств в Интернете.

Знаменитая формула Б. Гейтса «информация на кончиках пальцев» (information at your fingertips).

Так, для поиска информации в Интернете служат различные классы поисковых средств:

- каталоги *(directories):*
- подборки ссылок *(bookmarks):*
- поисковые машины *(search engines):*
- базы данных адресов электронной почты и т.д.

Каталог веб-ресурсов – постоянно обновляемая и пополняемая система ссылок на ресурсы, распределенные по иерархической структуре категорий.

На верхнем уровне каталога представлены самые общие категории (рубрики), например «наука», «бизнес», «развлечения » и т.д.

На нижележащих уровнях рубрики имеют более частный характер. Например, рубрика «наука» может делиться на категории «точные науки», «естественные науки» и «гуманитарные науки», последние – на философию, социологию, психологию, педагогику и т.д.

Русскоязычный каталог сайтов можно найти по адресу www.ru.

Коллекция ссылок представляет собой еще один способ организации информации во Всемирной паутине. Такая коллекция обычно составляется специалистом в определенной теме, постоянно обновляется и не содержит ненужной информации. Печатный аналог такой коллекции ссылок по использованию информационных технологий в лингвистике можно найти после библиографического списка. Некоторые примеры коллекций ссылок по обучению английскому языку приводит С.В. Титова.

Поисковые машины (или поисковые системы) – это специальные веб-страницы, позволяющие находить веб-ресурсы, текстовое содержание которых соответствует запросу пользователя.
В Международном каталоге поисковых машин (www.searchenginecolossus.com) зарегистрировано свыше 2300 поисковых систем из 232 стран.

К наиболее известным *поисковым машинам* относятся:

- ✔ AltaVista (www.altavista.com);
- ✔ Excite (www.excite.com);
- ✔ Yahoo! (www.yahoo.com);
- ✔ AOL (http://search.aol.com);
- ✔ MSN (http://search.msn.com);
- ✔ Google (www.google.ru);
- ✔ Яndex (www.yandex.ru);
- ✔ Rambler (www.rambler.ru);
- ✔ Апорт (www.aport.ru).

3. Характеристика поисковой системы Интернета. Информационно-поисковый язык.

Информационно-поисковый язык (ИПЯ) – формальный язык, предназначенный для описания содержания документов, хранящихся в ИПС, и запроса. Информационно-поисковые языки представляют собой знаковые системы со своим алфавитом, лексикой, грамматикой и правилами пользования. О специфике ИПЯ каждой поисковой системы, особенно о его «синтаксисе» (т.е. о правилах сочетания ключевых слов, вводимых в строку поиска) можно узнать на отдельных вкладках соответствующей поисковой системы. Например, в Яндекс такая вкладка называется «Помощь – Как искать».

Процедура описания документа на ИПЯ называется ***индексированием***. В результате индексирования каждому документу приписывается его формальное описание – ***поисковый образ документа***.
Аналогичным образом индексируется и запрос, которому приписывается поисковый образ запроса или ***поисковое предписание***.
Алгоритмы информационного поиска основаны на сравнении поискового предписания с поисковым образом запроса.

Степень соответствия документа запросу задается категорией ***релевантности***. При этом в процессе информационного поиска можно получить в выдаче значительный ***информационный шум*** – множество документов, формально релевантных, но не являющихся релевантными по смыслу.
Чтобы получить меньше информационного шума, пользователю следует уточнять свой запрос, используя для этого дополнительные настройки поисковой системы. Так, в *Google,* нажав вкладку «Расширенный поиск», можно задать поиск целых словосочетаний (а не отдельных составляющих их слов), ограничить язык выдачи, дату создания документа, часть документа, в которой используется слово, формат документа и т.д. Такие манипуляции увеличивают вероятность нахождения нужной информации уже в самом начале выдаваемого списка.

Результаты поиска могут характеризоваться с двух точек зрения: полноты и точности. ***Полнотой поиска*** (англ. *Recall)* называется мера, вычисляемая как отношение количества выданных релевантных документов к общему числу релевантных документов, содержащихся в информационном массиве. ***Точность поиска*** (англ. *Precision)* – это отношение количества выданных релевантных документов к общему числу документов в выдаче. Составить представление о полноте и точности поиска можно, сравнивая выдачи разных поисковых систем. При четком определении ключевых слов запроса и их синтаксической связи значения полноты и точности поиска будут стремиться к единице, т.е. к минимуму релевантных документов, что облегчает выбор человеком нужного результата поиска.

СПАСИБО ЗА ВНИМАНИЕ!