Из цикла лекций **«Современные Internet-технологии»** для студентов 5-го курса кафедры Компьютерных технологий физического факультета Донецкого национального университета

Поиск в Internet

ДонНУ, кафедра КТ, проф. В. К. Толстых

Содержание

Для реализации поисковых алгоритмов, технологий и средств взаимодействия поисковых систем с человеком сегодня интенсивно разрабатываются и внедряются интеллектуальные агенты.

Как работают каталоги

Поисковые узлы **каталоги** обслуживает большое количество людей (~100):

- **Классификаторы** – разрабатывают и совершенствуют рубрики своей информационной базы для Internet-документов,
- **Систематизаторы** – читают Internet-документы и, зная рубрики классификаторов, приписывают им классификационные индексы.

При классификации и систематизации информации здесь постоянно присутствует «человеческий» фактор.

Достоинства каталогов – простой доступ пользователей к популярной и качественной информации.

Недостатки – любая оценка документа классификатором и систематизатором является социальным действием, она связана с их культурой, мировоззрением, глубиной и широтой знаний.

Как работают ИПС

Интеллектуальные агенты ИПС – это комплекс программ:

- **Spider** («паук») — программа, которая загружает в поисковую машину Web-страницы. Работает аналогично браузеру, но ничего не отображает ни на каком экране.
- **Crawler** («червяк», или «путешествующий паук») — программа, способная найти на Web-странице все ссылки на другие страницы. Ее задача — определить, куда дальше должен ползти «паук», руководствуясь ссылками или заранее заданным списком адресов.
- **Indexer** (индексатор) — программа, которая «разбирает» страницу на составные части и анализирует их. Вычленяются и анализируются заголовки Web-страниц, заголовки в документах, ссылки, текст документов, его выделения…
- **Database** (база данных) — хранилище данных в виде инвертированного индекса, где для каждого слова из страниц доставленных пауком перечислены все места (URL документов, позиция слова, цвет и размер шрифта...), в которых слово встретилось.
- **Search Engine Results Engine** (система выдачи результатов поиска) решает, какие страницы удовлетворяют запросу пользователя и в какой степени. Именно с этой частью поисковой системы «общается» пользователь. Это агент с наибольшим интеллектом.

Полнота

Два аспекта: **полнота охвата** , **полнота отклика**

Полнота *охвата* – это общее количество проиндексированных из Internet документов.
Полнота *отклика* определяется по формуле:

$$П = \frac{N_1}{N} 100\% ,$$

где N_1 – количество полученных документов, N – количество имеющихся в базе документов формально соответствующих запросу. В идеале должно быть 100%.

Полнота тесно связана с **оперативностью обновления** информации.

Два аспекта полноты

Релевантность

Релевантность – соответствие полученной информации отправленному запросу:

$$П = \frac{N_1}{N} 100\%,$$

где N_2 – количество документов, ***соответствующих*** запросу, N – общий объём полученной информации.
В идеале должно быть 100%.

Механизмы расчёта релевантности

Пертинентность – соотношение полезной для пользователя информации N_3 к общему объёму полученной информации:

$$П = \frac{N_1}{N} 100\%,$$

Средства повышения пертинентности:

1. уточнение формулировки запроса,
2. ранжирование документов по весовым коэффициентам,
3. Внедрение интеллектуальных технологий поиска.

Релевантность и пертинентность

Лидеры ИПС

Международные:

- **http://www.google.com**
- **http://www.bing.com**
- **http://search.yahoo.com**
- **http://www.ask.com**
- **http://www.alltheweb.com**
- **http://www.lycos.com**
- www.go.com

Российские:

- **http://www.yandex.ru**
- **http://www.rambler.ru**
- **http://www.aport.ru**

Украинские:

- **http://meta.ua**
- **http://uaport.net**

Основные логические операторы

Оператор	Яndex	Google
Логическое И	& пробел (в пределах предложения) && (в пределах документа)	пробел
Логическое ИЛИ	|	OR
Логическое НЕ	~ (в пределах предложения) ~~ или - (в пределах документа)	–
Группировка Приоритет операций: NOT, AND, OR	()	()

Примеры профессиональных запросов к ИПС

Запрос к системе "Интегрум" по теме"Услуги связи:

"услуги связи" или "междугородные переговоры" или "телефонные переговоры" или "мобильная связь" или "фиксированная связь" или "сотовая связь" или "сотовый оператор" или "средства связи" или "телефонная связь" или "спутниковая связь" или "космическая связь" или GPRS или ростелеком или связьинвест или госкомсвязь или госкомтелеком или госсвязьнадзор или телекоммуникации или электросвязь или АТС или ГТС или минсвязи или "министерство связи" или "волоконно-оптическая линия связи" или ВОЛС

Запрос к системе InfoStream по теме "Мобильная связь":

(((мобильн~связ) | (мобільн~зв'яз) | (сотов~связ) | (стільник~зв'яз) | (беспроводн~связ) | (бездрот~зв'яз) | (бесперебойн~связ) | (безперебійн~зв'яз) | j2me]| ems]| 3g]| gprs]| ggsn]| sgsn]| sms]| mms]| ems]| bluetooth]| mms]| tdma]| multipoint]| pcs]| cdma]| ofdm]| vpn]| wap]| umts]| gsm)&((моб~телефон)| (стільник~телефон)| (сотов~телефон))) ! this.is

PageRank

– алгоритм оценки популярности Интернет-страниц.

Сергей Брин в 1998 году предложил идею:

определять рейтинг страницы через количество ведущих на неё ссылок и рейтинг ссылающихся страниц.

SEO: Search Engine Optimization

Поисковая оптимизация направлена на увеличение количества посетителей Web-сайта за счёт повышения рейтинга страниц сайта (без оплаты поисковым компаниям).

Факторы, влияющие на поисковый ранг

- <title>, <h1-6> - должны быть достоверными с нужными ключевыми словами
- имена каталогов, файлов должны быть «ключевыми словами». Отдельные слова в имени файла страницы должны отделяться «-», а не «_», т. к. «-» ИПС трактуют как пробел и индексируют все слова, а «_» - как объединение слов.
- ссылки (отсутствие) на страницы спама или «дурного общества»
- чем старее домен, страница (при этом активно изменяется), тем выше рейтинг
- ссылки со «старых» сторонних сайтов повышают рейтинг страницы
- длительные (более года) оплаты за домен повышают рейтинг сайта (спамеры покупают домены не более чем на год)
- количество, качество и релевантность входящих ссылок.
- GET-параметры исходящих ссылок индексируются вместе со ссылками, поэтому они должны иметь осмысленные, постоянные ключевые слова
- текст, окружающий ссылки, должен быть семантически родственным. Это повышает ранг ссылки и страницы на которую она ссылается
- ссылка на страницу с множеством исходящих ссылок понижает ранг ссылки
- ссылки между страницами из IP одного класса C понижаются в рейтинге, т. к. похожи на механизм искусственного рейтинга
- ссылки из доменов .edu, .gov имеют повышенный ранг
- важные ссылки на страницах не должны располагаться в конце страницы

Штрафование поискового ранга

- ссылки с разными GET-запросами, приводящими к одной и той же странице. Нельзя в GET вставлять параметры сеансов, т. к. они изменяются
- перенаправление на др. страницы на клиенте считается спамом
- перемещение или изменение имени страницы снижает её рейтинг
- страницы с дублированным контентом:
  - вследствие архитектуры сайта (в т. ч. страницы для печати, одинаковые <meta>, <title>…)
  - вследствие кражи контента
  - нельзя в ссылках указывать имена файлов, загружаемых по умолчанию, т. к. такие файлы будет индексироваться дважды

Преодолеть штрафные баллы за дублирование контента можно через закрытие соответствующего контента от индексирования поисковиками. Для этого надо поместить в корень сайта файл

robots.txt :

User-agent: * *- для всех типов поисковиков*

Disallow: /admin/ *- для всех URL, начинающихся с* /admin/

Disallow: /*Intra/ *- для всех URL, содержащих где-либо* /Intra/

Disallow: /file.txt *- для всех URL, начинающихся с* /file.txt/