Поиск информации.

Борисов В.А.
КАСК – филиал ФГБОУ ВПО РАНХ и ГС
Красноармейск 2011 г.

Поиск информации

- Задача, которую человечество решает уже многие столетия.

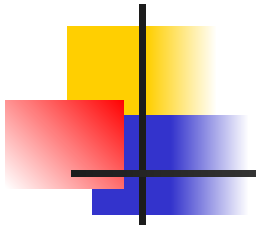
- Все найденные за много лет средства и приемы поиска информации доступны и эффективны и при поиске информации в Интернет.

“ПЕРТИНЕНТНЫЙ ДОКУМЕНТ”

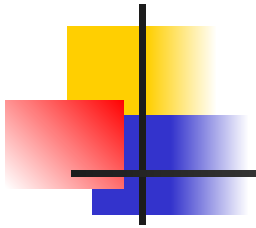
- Слово “пертинентный” происходит от английского “pertinent”, что значит “относящийся к делу, подходящий по сути”.

Цель информационного поиска

- Найти все пертинентные и только пертинентные документы (мы хотим найти “только то, что хотим, и ничего больше”).
- Эта цель - идеальна и пока недостижима.



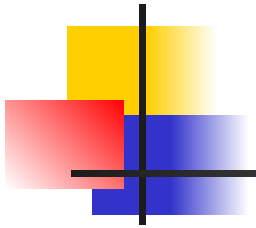
- Для того, чтобы было с чем сравнивать, необходимо некоторое количество непертинентных документов.
- Эти документы называются - “ШУМ”.



- Когда документов много, используется информационно-поисковая система (ИПС).
- В этом случае информационная потребность должна быть выражена средствами, которые “понимает” ИПС - должен быть сформулирован ЗАПРОС.

РЕЛЕВАНТНОСТЬ

- Степень соответствия документа запросу.



Виды информационно поисковых систем

Классификационные ИПС

- В классификационных ИПС используется иерархическая (древовидная) организация информации, которая называется КЛАССИФИКАТОРОМ.

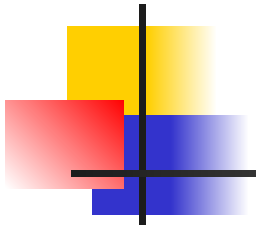
- Разделы классификатора называются РУБРИКАМИ.
- Библиотечный аналог классификационной ИПС - систематический каталог.

Предметная ИПС Web-кольца

- Поиск названия нужного предмета своего интереса (предметом может быть и нечто невещественное, например, индийская музыка), а с названием связаны списки соответствующих ресурсов Интернет.

Словарные ИПС

- Основная идея словарной ИПС - создать словарь из слов, встречающихся в документах Интернет, в котором при каждом слове будет храниться список документов, из которых взято данное слово.



- Два основных алгоритма работы словарных ИПС:

  -с использованием ключевых слов,

  -с использованием дескрипторов.

Использование ключевых слов

- Для оценки содержимого документа используются только те слова, которые в нем встречаются, и по запросу ИПС сопоставляет слова из запроса со словами документа, определяя по количеству, расположению, весу слов из запроса в документе его релевантность.

Использование дескрипторов

- Индексируемые документы переводятся на некоторый дескрипторный информационный язык.
- Дескрипторный информационный язык, как и любой другой язык, состоит из алфавита (символов), слов, средств выражения парадигматических и синтагматических отношений между словами.

Ранжирование результатов поиска

- Все ИПС в настоящее время уделяют основное внимание именно алгоритму ранжирования полученных ссылок.

Критерии при ранжировании в ИПС

- наличие слов из запроса в документе, их количество, близость к началу документа, близость друг к другу;
- наличие слов из запроса в заголовках и подзаголовках документов;
- количество ссылок на данный документ с других документов;
- «респектабельность» ссылающихся документов.

Современные проблемы поисковых систем

- Когда эти технологии разрабатывались никто из разработчиков не представлял себе, что Интернет станет глобальной информационной средой.

Архитектура

- crawler (сборщик) - осуществляет сканирование Интернет ресурсов в поисках изменений на страницах;
- indexer (индексатор) - индексирует ресурсы, строит базы данных по ключевым словам, хранит эти базы данных в виде, удобном для поиска по ним;
- gateway (шлюз) - осуществляет прием запросов от пользователей и выдачу им информации из базы данных.

Алгоритмы поиска и ранжирования

- Основной проблемой современных поисковых систем является то, что по причине фактически устаревшей архитектуры они не могут обеспечить качественный поиск информации.

Основные моменты новой архитектуры ИПС

- Переход к распределенной модели вычислений;
- Переход от модели «один поиск на всех» к модели персонального поиска;
- Переход от критериев релевантности к критерию пертинентности;
- Переход от поиска только текстовой информации к распознаванию и поиску мультимедийной информации.