Поисковые системы

Причины и предпосылки

1. Появление WWW
2. Графические браузеры
3. Стремительный рост объемов разнородных документов в Интернете.
4. Сложность определения местоположения получения информации.

- До WWW служба поиска
  - По ftp-серверам Archi
  - По gopher-серверам Veronica

Основные типы

- Поисковые машины
- Тематические каталоги

Тематический каталог

- В www:
  - сперва появлялись личные коллекции ссылок.
  - Развитие – тематические каталоги (www.yahoo.com 1994 год).
- Составные части:
  - Иерархическая совокупность тем (рубрик)
    - (обычно постоянно совершенствуется и использует перекрестные ссылки)
  - База описаний ресурсов с привязкой к рубрикам (возможно к нескольким)
  - Механизм отслеживания описания ресурсов и пополнения базы.

Функционирование тематических каталогов требует значительной «ручной» работы операторов.

Поисковая машина

- Появились с популярностью www
  - Одни из первых (1993 год)
    - ALIWEB (просматривала META теги)
    - Exite (анализировала статистику появления слов в документе)
    - Lycos (индексировал страницу целиком)
- Поисковая машина состоит из
  - программы-паука (робот), которая просматривает сайты Интернета и индексирует их в автоматическом режиме.
  - базы данных (индекса), в которой находится информация о просмотренных сайтах.
    - В современных системах база содержит и сами документы (размером до нескольких десятков килобайт).

Основная часть работы выполняется в автоматическом режиме.

Языки запросов

- Запрос – обращение клиента к поисковой машине, составленное на языке запросов.
- Выражение на языке запросов состоит из «ключевых слов», операторов булевой алгебры и других служебных символов.
- Каждая поисковая машина имеет свой язык запросов

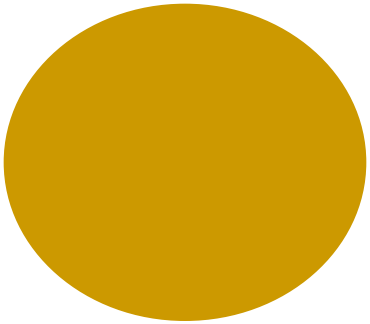
- Пример операторов булевой алгебры:

И	ИЛИ	НЕ
AND	OR	NOT
+	*пробел*	-

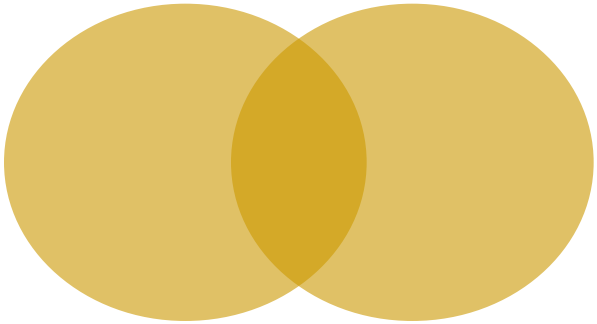
- Операторы могут группироваться с помощью скобок

Логические (булевы) выражения

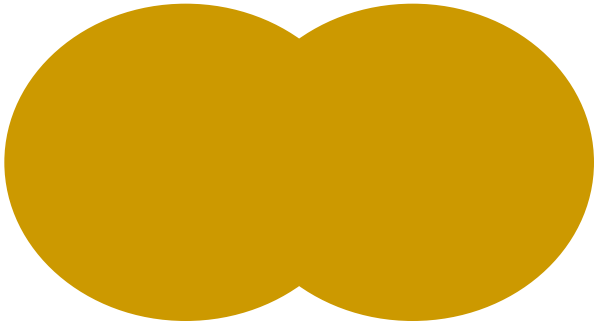
A



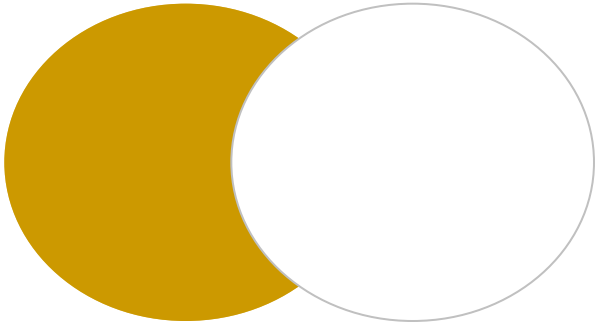
A AND B



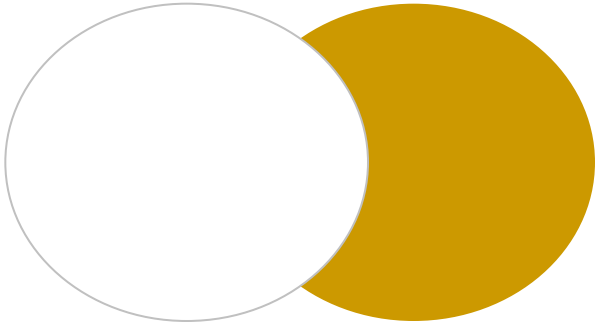
A OR B



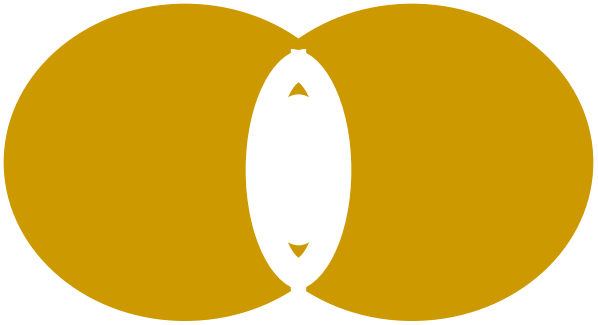
A NOT B



B NOT A



A OR B NOT (A AND B)



Пример запросов в ПМ Яндекс

- 1. Найти документы с любым словом из двух:
  Слово1 | *Слово2*
- 2. Найти документы с двумя словами:
  Слово1 && Слово2
- 3. Найти документы с двумя словами в одном предложении:
  Слово1 & Слово2
- 4. Найти документы, которые содержат первое слово, но не содержат второго: *Слово1 ~~ Слово2*
- 5. По умолчанию Яндекс ищет с использованием морфологии, можно отключить:
  !Слово
  Т.е. будут найдены документы с указанной словоформой.
- 6 …

Сравнение

тематические каталоги

поисковые машины

- Тематические каталоги
  - + относительно точное соответствие ресурса заявленной теме.
  - + «тщательно» отобранные названия тем.
  - - медленность обновления
  - - небольшой объем базы
- Поисковые машины
  - + Большая скорость обновления
  - + Большой объем базы
  - - Возможно присутствие большого количество «мусора» в результатах поиска

В большинстве случаев поисковые системы включают в себя и тематические каталоги и поисковые машины (или пользуются «чужими»)

Поисковые системы в Рунете

- Yandex
- Mail (использует поисковый механизм Yandex с декабря 2005)
- Rambler
- Aport
- Другие:
  - Punto
  - …

Крупнейшие поисковые ресурсы (зарубежные)

- Yahoo
- Google
- MSN Search
- Ask Jeeves – Европа США (www.ask.com) (поиск на естественном языке 1997)

- Исторически известные:
  - Excite (принадлежит Ask Jeeves)
  - Lycos – один из первых поисковиков
  - Altavista (принадлежит Yahoo)

Google

- Студенты кафедры информатики Стэндфордского университета: Ларри Пейдж и Сергей Брин разработали поисковый "движок" BackRub (анализ обратных ссылок)
- 1996 начало, 1998 – запуск
- Первая контекстная реклама
- Параллельные проекты (карта мира, изображения земли, луны, марса, оцифровка книг крупнейших библиотек, Web-приложения)

Специализированные поисковики

- Только по определенным ресурсам (mp3)
- Только по ftp ресурсам (www.filesearch.ru)
- По какой-то тематике (обычно каталоги), например медицинской.
- Метапоисковики (www.metabot.ru)
- www.dogpile.com