ПРИНЦИП РАБОТЫ ПОИСКОВОЙ СИСТЕМЫ

Гурова Марина БСТ-212

Поисковая система (англ.*search engine*) — это компьютерная система, предназначенная для поиска информации. Одно из наиболее известных применений поисковых систем — веб-сервисы для поиска текстовой или графической информации во Всемирной паутине. А также это аппаратно-программный комплекс, который предназначен для осуществления функции поиска в интернете, и реагирующий на пользовательский запрос который обычно задают в виде какой-либо текстовой фразы (или точнее поискового запроса), выдачей ссылочного списка на информационные источники, осуществляющейся по релевантности. Самые распространенные и крупные системы поиска: Google, Bing, Yahoo, Baidu. В Рунете – Яндекс, Mail.Ru, Рамблер.

Запрос в поисковой системе

- Для поиска информации с помощью поисковой системы пользователь формулирует поисковый запрос. Работа поисковой системы заключается в том, чтобы по запросу пользователя найти документы, содержащие либо указанные ключевые слова, либо слова, как-либо связанные с ключевыми словами. При этом поисковая система генерирует страницу результатов поиска. Такая поисковая выдача может содержать различные типы результатов, например: веб-страницы, изображения, аудиофайлы. Некоторые поисковые системы также извлекают информацию из подходящих баз данных и каталогов ресурсов в Интернете.

Методы поиска

- По методам поиска и обслуживания разделяют четыре типа поисковых систем: системы, использующие поисковых роботов, системы, управляемые человеком, гибридные системы и мета-системы. В архитектуру поисковой системы обычно входят:
- поисковый робот, собирающий информацию с сайтов сети Интернет или из других документов
- индексатор, обеспечивающий быстрый поиск по накопленной информации, и
- поисковик — графический интерфейс для работы пользователя.

На практике обычно поступают следующим образом. Изначально поступившую информацию оценивают с точки зрения релевантности. Если информация релевантна — вопрос в ее достоверности. Затем — в ее актуальности. А после этого при необходимости осуществляется оценка по иным критериям. Часто для ускорения процесса оценки используют упрощенный набор критериев.

Принцип работы поисковых систем

- Поисковые системы работают, храня информацию о многих веб-страницах, которые они получают из HTML страниц. Поисковый робот или «краулер» — программа, которая автоматически проходит по всем ссылкам, найденным на странице, и выделяет их.
- Поисковая система анализирует содержание каждой страницы для дальнейшего индексирования. Слова могут быть извлечены из заголовков, текста страницы или специальных полей — метатегов. Индексатор — это модуль, который анализирует страницу, предварительно разбив её на части, применяя собственные лексические и морфологические алгоритмы.
- Как правило, системы работают поэтапно. Сначала поисковый робот получает контент, затем индексатор генерирует доступный для поиска индекс, и наконец, поисковик обеспечивает функциональность для поиска индексируемых данных. Чтобы обновить поисковую систему, этот цикл индексации выполняется повторно.

История развития поисковых систем

- На раннем этапе развития сети Интернет Тим Бернерс-Ли поддерживал список веб-серверов, размещённый на сайте ЦЕРН. Первой компьютерной программой для поиска в Интернете была программа Арчи. Она была создана в 1990 году, студентами, изучающими информатику в университете Макгилла в Монреале. Программа скачивала списки всех файлов со всех доступных анонимных FTP-серверов и строила базу данных, в которой можно было выполнять поиск по именам файлов. Однако, программа Арчи не индексировала содержимое этих файлов, так как объём данных был настолько мал, что всё можно было легко найти вручную.