Принципы работы поисковых систем

Гурова Марина БСТ-212

Общий принцип работы любой поисковой системы условно можно разделить на следующие этапы:

- **Сбор информации.** Специальная программа сканирует веб-пространство, открывает каждый доступный для неё сайт и анализирует его по заданным алгоритмам.
- Все **документы закачиваются на сервер поисковой системы** и создается база данных, которая содержит информацию о сайте.
- На основе полученных по сайту данных проводится построения **индекса.** То есть определяется, какие данные на нём содержатся, к какой группе запросов относятся данный контент их можно отнести и так далее.
- **Программа определяет релевантность страницы**, в момент когда она получает пользовательский поисковый запрос, на его основе предоставляет перечень сайтов, которые по результатам индексирования содержат запрашиваемую информацию.
- **Сервис проводит ранжирование** результатов выдачи. То есть выстраивает порядок ссылок, которые будут показаны пользователю, отправившему запрос.

Описанный принцип работы информационно поисковых систем — это лишь условное пояснение, как работает тот же Google или Яндекс. Но вот алгоритмы, которые они используют для обхода, сайтов, индексации и ранжирования, обычным пользователям неизвестны, каждая поисковая система применяет свои алгоритмы и постоянно их совершенствует, так как обработка информации занимает большое количество ресурсов сервера, расходы на который лежат на поисковой системе.

Поисковые системы бывают нескольких подвидов и существуют и другие вариации таких сервисов:

- управляемые человеком (то есть каталог сайтов, каждый пункт которого и общая их база данных сформированы вручную пользователем);
- гибридные поисковые системы (где часть работы выполняет человек, часть — программ);
- мета-системы (которые не составляют базу данных, а дают результат сразу из нескольких поисковых сервисов).

Принципы работы поисковой системы

Главные этапы составления базы данных для поисковых сервисов — это индексация и ранжирование сайтов. И чтобы результативность итоговой выдачи была точной, сейчас применяется схема машинного обучения. То есть поисковику демонстрируют для сравнения 2 противоположных результата и указывают, по какой схеме необходимо выполнять их ранжирование. Таким образом система понимает, какой сайт «полезный», какой — «менее полезный».

Всё это позволяет вывести отдельный индекс — релевантность (условно можно назвать «рейтингом»). Он присваивается каждому сайту, представлен в виде дробного числа. Чем выше релевантность — тем выше будет позиция ресурса в выдаче на запрос пользователя. Это — основные принципы работы поисковых систем, используемых сегодня. И этот процесс тоже состоит из нескольких этапов.

Сбор данных

После создания сайта и получения на него ссылки, система автоматически анализирует его с помощью инструментов Spyder и Crawling. Информация собирается и систематизируется из каждой страницы.

Индексация

Индексация выполняется с определенной периодичностью. И по её прохождению сайт добавляется в общий каталог поисковой системы. Результата этого процесса — создание файла индекса, который используется для быстрого нахождения запрашиваемой информации на ресурсе.

Обработка информации

Система получает пользовательский запрос, анализирует его. Определяются ключевые слова, которые в дальнейшем и используются для поиска по файлам индекса. Из базы данных извлекаются все документы, схожие на пользовательский запрос.

Ранжирование

Из всех документов, отобранных для выдачи, составляется список, где каждому сайту отведена своя позиция. Выполняется на основании ранее вычисленных показателей релевантности.

На этом этапе принцип работы поисковых систем немного разнится. Формула ранжирования — тоже уникальная. Но ключевые факторы, влияющие на релевантность сайта, следующие:

- индекс цитируемости (как часто сторонние ресурсы ссылаются на информацию из конкретной страницы);
- авторитетность домена (определяется на основании его истории изменения);
- релевантность текстовой информации по запросу;
- релевантность иных форматов контента, представленных на странице;
- качество оптимизации сайта.