Имеющиеся в 90-х годах поисковики с трудом справлялись со своей задачей. Результаты поисковой выдачи имели очень низкую корреляцию с тем, что хотел увидеть в ответ на свой запрос пользователь. Дело в том, что тогда основным маркером (фактором), по которому осуществлялось определение релевантности и ранжирование документов в выдаче, была частота использования слов из запроса пользователя в документе. Но такой критерий отбора очень легко поддается накрутке со стороны веб-мастеров простым увеличением частоты использования ключевых слов.

Один из создателей Google - Ларри Пейдж, с детства на примере своих родителей, вращавшихся в научных кругах, видел и понимал, что авторитет того или иного ученого во многом зависит от того в скольких научных работах на него, ссылаются, как на первоисточник или как на авторитетного специалиста. Однажды у него возникла идея использовать подобную систему ранжирования для поиска в интернете. Так в результате появился всем известный фактор ранжирования, который учитывается поисковиками до сих пор — PageRank.

PageRank поистине совершил революцию и позволил поднять качество поиска будущего поисковика Google на недосягаемую высоту. PageRank позволял учитывать при ранжировании документов в будущем поисковике Google не только количество, но и качество ведущих на ту или иную веб-страницу.

Google был запущен в 1998 году выпускниками Стэндфордского университета Сергеем Брином (Sergey Brin) и Ларри Пейджем (Larry Page), в свое время работавшими над учебным проектом по идентификации смысловых элементов в структуре Web-ссылок. Они были поражены огромным значением так называемых «обратных ссылок» (то есть страниц, ссылающихся на сайт) и поняли, что их можно использовать для того, чтобы создать более эффективную поисковую систему.

Сначала поисковая система называлась “Googol”, что означало число 10, возведенное в степень 100 (единичка со ста нолями). Это подчеркивало бесконечное число документов в сети Интернет. Однако, после представления проекта своему первоначальному инвестору, Брин и Пейдж получили чек на имя “Google” и для того чтобы получить деньги, им пришлось изменить название с “Googol” на “Google”.

Google заставил мир поисковых систем перевернуться с ног на голову благодаря своей концепции PageRank, которая оказалась настоящим технологическим прорывом и которую сейчас использует большинство ведущих поисковых систем для обеспечения более качественного поиска. Технология поиска PageRank работает путем установления структуры ссылок во всей сети, а затем ранжирует каждую отдельную страницу, основываясь на числе и значимости ссылок на нее на других страницах.

Это поисковая система Google быстро приобрела известность как предоставляющая беспрецедентно хорошие результаты.

Официальный запуск Google состоялся в сентябре 1999 года (до этого действовала лишь предварительная бета-версия сайта). Популярность Google стремительно росла. В 2000 году на долю Google приходились около 5% поисковых запросов в интернете, в 2003 году - 32%, а к 2004 году через него проходило уже более половины всех запросов. В 2000 году у поисковика появились версии интерфейса на ряде языков помимо английского. К 2008 году Google был доступен на 116 языках, в том числе на некоторых искусственных и вымерших языках.

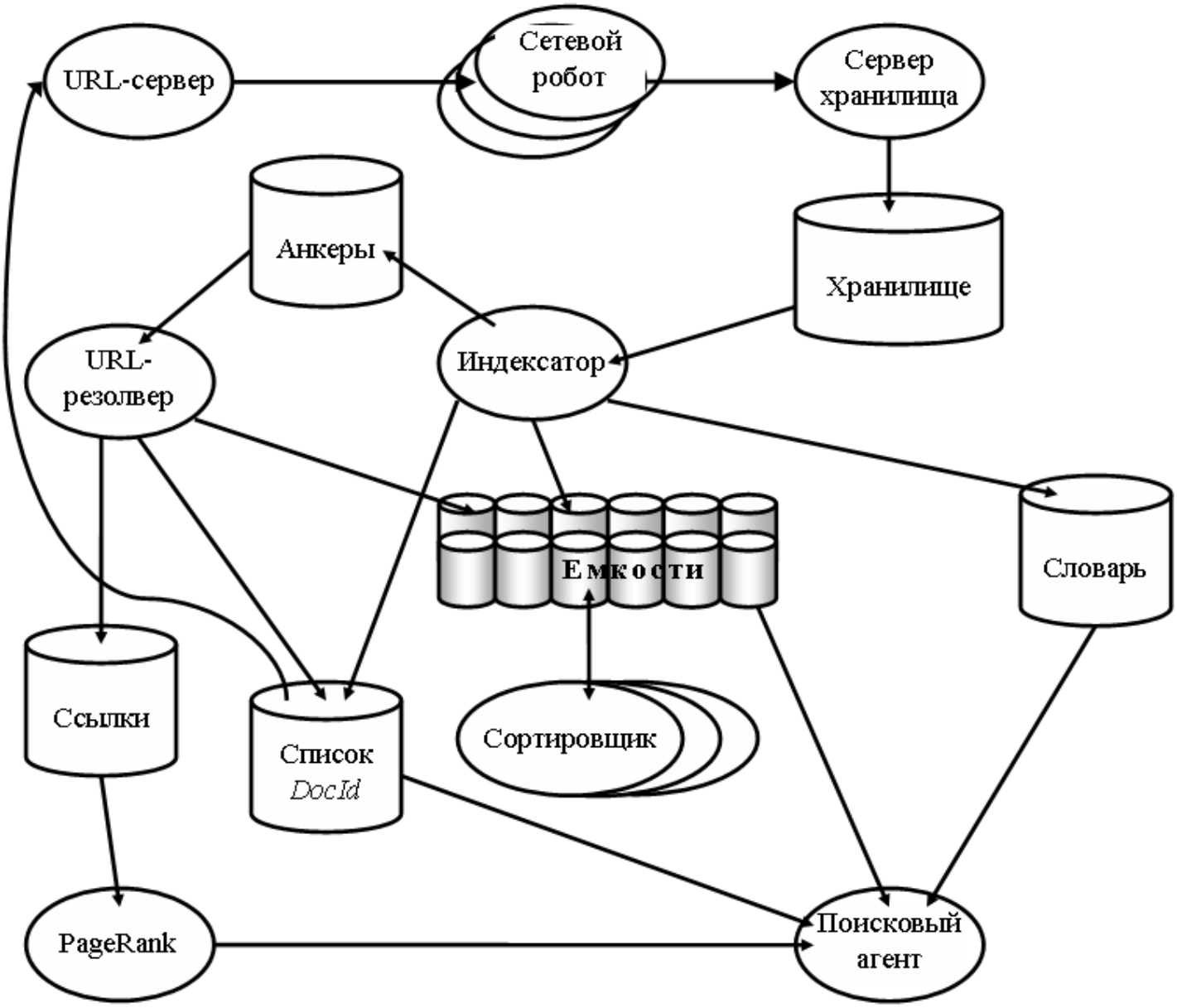
Архитектура инфомационно-поисковой системы Google

В любой поисковой системе можно выделить три базовых части:

1. Робот (краулер, спайдер, индексатор). Робот отвечает за сбор информации. То есть робот эмулирует работу пользователя, загружая страницы и сохраняя их в базе данных.

2. База данных. В базе данных хранится и сортируется собранная роботом информация.

3. Клиент. В этой части обрабатываются пользовательские запросы. В действительности клиент может быть разнесен по нескольким физически несвязанным компьютерам. Однако, стоит отметить, что все эти компьютеры должны иметь доступ к базе данных.



1. URL сервер (URL Server) - список всех адресов

2. Сетевой робот (Crawler) – робот («паук»), который загружает страницы из списка адресов и передает в Сервер хранилища.

3. Сервер хранилища (Store Server) сохраняет страницы в Хранилище (Repository), чаще всего в виде HTML документа. При этом вся дополнительная информация, такая как картинки, flash-анимация и прочее, не сохраняется.

4. Индексатор (Indexer) разбирает сохраненные в Хранилище (Repository) HTML-документы в последовательности слов и сохраняет их в Базы данных (Barrles).

5. Словарь (Lexicon) - список всех слов. Чаще всего слова храняться в таблице с двумя полями “номер” и “слово”. Таким образом, достигается экономия места в базе данных, так как длинные слова заменяются достаточно коротким номером

6. Анкеры (Anchors) выделенные Индексатором (Indexer) ссылки (URL).

7. URL Resolver - обработчик URL. Если находятся новые ссылки, то они передаются в URL сервер

8. Ссылки (Links) - определяет какие сайты на какие ссылаются и передает эту информацию в PageRank.

9. PageRank определяет рейтинг сайта, основным критерием является количество ссылок на этот сайт

10. Поисковый агент (Searcher) - клиент. Чаще всего клиент пользуется статической базой данных, которая обновляется примерно раз в сутки.

Сетевой робот, используемый Google, имеет централизованную архитектуру. Такая архитектура состоит из нескольких потенциально распределенных конкурентных компонентов, имеющих центральный пункт синхронизации (например, очередь задач или специальный компонент-координатор).

Поисковая система Google использует в своей поисковой системе три поисковых робота - бота - Freshbot, DeepCrawl, Adsensebot.

Freshbot – этот поисковый робот - бот заходит на сайт, сканирует его, определяя наиболее популярные странички, и индексирует их. Поисковый робот - бот Freshbot посещает сайты в среднем 1 раз в два дня, но посещаемость его резко возрастает если сайт популярен, имеет хорошую посещаемость и довольно часто обновляется. Например, такие сайты как myspace.com или youtube.com он посещает каждые 5-10 минут. Еще одной из задач поискового робота Freshbot является сканирование всех страниц сайта для сбора всех ссылок в базу данных, после чего эта база передается другому роботу - боту DeepCrawl, который будет сканировать все эти ссылки.

DeepCrawl - этот поисковый робот - бот, который получив базу данных с ссылками сайта от поискового робота - бота Freshbot, приступает к сканированию этих ссылок, добавляя их в свой индекс. DeepCrawl посещает сайты всего 1 раз в месяц, поэтому результаты индексации этого поискового робота приходится ждать долго.

Adsensebot - поисковый робот – бот, предназначенный для страниц сайта, которые транслируют контекстную рекламу Adsense. Если Вы установить на свои странички сайта скрипт от Google Adsense, скрипт Adsense посылает команду для Adsensebot, после чего этот поисковый робот приходит на сайт и сканирует страницы для более точного определения релевантности объявлений по отношению к контенту странички. Например, если на страничке сайта пишется о машине, то поисковый робот определит наиболее частое употребление ключевых слов, связанное с машиной, и естественно предоставит свою рекламу той же тематики

Сравнение поисковых систем яндекс и google

Алгоритмы Яндекса и Google различаются кардинально. Яндекс строит свои позиции исходя из уникальности текстов на сайтах, а Google - исходя из количества ссылок, которые ведут на рассматриваемый сайт.

За многие годы работы с этими поисковыми системами, многие пользователи сделали для себя вывод, что с Google работать гораздо проще. Плохие с точки зрения посетителей сайты он быстро вычисляет, накладывает на них самые разнообразные штрафные санкции (фильтры), а то и вовсе исключает их из поисковой выдачи (отправляет в бан).

Даже если количество ключевых фраз велико, но Google видит, что посетители на этом сайте "сидят" довольно продолжительное время, то этот положительный фактор перекрывает многие отрицательные. Google готов даже «закрыть глаза» на обман своих поисковых роботов, если посетителю сайт нравится.

И Google, и Яндекс предоставляют пользователю возможность искать документ, в котором не содержится определенного слова; документ в котором присутствует любое из слов запроса, и документ, в котором встречается абсолютно точное вхождение запроса. В каждой из поисковых систем за это отвечают различные операторы. Примеры использования поисковых операторов представлены в таблице 1.

Уникальными для Яндекса являются операторы:

1. /N, в котором N заменяется на число, обозначающее количество слов, которое может разделять в документе слова запроса;
2. ! осуществляет поиск без учета морфологии запроса;
3. & и && осуществляют поиск слов, встречающихся в одном предложении и на одной странице соответственно.

Примеры использования поисковых операторов.

Действие	Яндекс	Google
Строго все слова запрос	анализ архитектура сервер	анализ архитектура сервер
Поиск документа, в котором не содержится слов после знака	анализ архитектуры ~ здание	анализ архитектуры - здание
Ищет любое из слов запроса	анализ архитектура сервер	анализ OR архитектура OR сервер
Ищет точное вхождение запроса	"анализ архитектуры сервера"	"анализ архитектуры сервера"
Замена любого слова	-	Google *
Числовой интервал поиска	-	Google 10..100
Слова запроса встречаются в одном предложении	архитектура & сервер	-
Слова запроса находятся на одной странице	архитектура && сервер	-
Слова на расстояние указанного числа слов	архитектура /2	-
Поиск без учета морфологии	сервер	-

Поисковый алгоритм Google имеет несколько существенных преимуществ:

1. Использование механизма PageRank, который отображает "важность" сайта и влияет на выдачу результатов поиска. PageRank очень похож на индекс цитирования у Яндекса (тоже зависит от количества и качества ссылок на ресурс). Но в отличие от Яндекса, влияние PageRank у Google не настолько значительно, поэтому люди в Google находят именно то, что и ищут.
2. Google ищет не только гипертекстовые файлы (html), но и файлы в формате PDF, DOC, PostScript, Corel, WordPerfect и др.
3. Поисковая система Google обладает также возможностью поиска изображений. При этом в запросе можно указать желаемый размер, глубину цвета, формат файла.
4. В отличии от многих поисковиков, роботы Google индексируют все страницы, а не только самые главные.
5. Все страницы Google кэширует (заносит в свою базу), и разрешает человеку, производящему поиск, смотреть документ, не открывая его в первоисточнике, а беря из кэша Google (что часто намного быстрее).
6. Google разрешает настроить каждому пользователю язык интерфейса поисковой машины, выбрать языковые зоны для поиска, количество сообщений при выдаче результатов и др.
7. Пользователи Microsoft Internet Explorer, Mozilla Firefox и Opera могут установить себе программу Google Toolbar, которая создает новую панель инструментов, позволяющую искать в Google, не заходя на сам сайт.

Заключение

Сегодня поисковая система Google – одна из крупнейших в мире. Миллионы пользователей Интернета во всех странах ежедневно пользуются Google, потому что он быстр и прост в использовании и имеет огромную базу данных. Но самый главный аргумент – это то, что он действительно работает, и можно легко найти все, что нужно. Google удалось добиться большой популярности за короткое время, благодаря принципиально новому подходу в поиске информации в Интернете.