ЭВОЛЮЦИЯ АЛГОРИТМОВ КРУПНЫХ ПОИСКОВЫХ СИСТЕМ И ПРИНЦИП РАНЖИРОВАНИЯ

EVOLUTION OF ALGORITHMS OF LARGE SEARCH ENGINES AND THE PRINCIPLE OF RANKING

УДК 025

Лымарь Владислав Андреевич, студент, Уральский государственный экономический университет, Россия, г. Екатеринбург

Елифанова Марина Александровна, студент, Уральский государственный экономический университет, Россия, г. Екатеринбург

Lymar Vladislav Andreevich, vlad.lymar@list.ru

Marina Epifanova, vlad.lymar@list.ru

Аннотация

Многие пользователи при поиске информации в интернете не задумываются, какие сложные вычислительные операции происходят в данный момент и какие усилия компании предпринимали с самого своего зарождения, чтобы просто вывести по запросу прогноз погоды. В данной статье есть подробное описание эволюции алгоритмов поиска и их принципов.

Ранжирование — сортировка сайтов в поисковой выдаче, применяемая в поисковых системах. Существует множество факторов для ранжирования, среди которых можно отметить рейтинг сайта, количество и качество внешних ссылок, релевантность текста к поисковому запросу, на основании которых поисковая система формирует список сайтов в поисковой выдаче.

Annotation

When searching for information on the Internet, many users do not think about what complex computing operations are taking place at the moment and what efforts the company has made since its inception to simply display the weather forecast on

request. This article provides a detailed description of the evolution of search algorithms and their principles.

Ranking — sorting of sites in search results used in search engines. There are many factors for ranking, including the site's rating, the number and quality of external links, and the relevance of the text to the search query, which are used by the search engine to create a list of sites in the search results.

Ключевые слова: Алгоритмы поиска, ранжирование, анализ данных, веб-аналитика, SEO.

Keywords: Search algorithms, ranking, data analysis, web Analytics, SEO.

Стремительное развитие сети интернет заставляет огромные корпорации адаптироваться к изменениям «информационной игры» 21 века. Многие считают, что интернет – главная проблема человечества. Действительно, существует огромное количество совершенно бесполезной информации, которая порой может не только заставить усомниться человека в своих знаниях, но и навредить, однако, у интернета есть и другая сторона, которая кардинально перевернула вектор развития человечества.

Если спросить жителя России, какую поисковую систему он использует, то в подавляющем большинстве случаев ответ будет «Яндекс» или «Google». Существуют и другие поисковые системы, но используются они крайне редко. По данным из открытого доступа статистики Liveinternet в 2017 и 2018 годах, более чем в 50% случаев жители России выбирали в качестве поисковика «Яндекс», за «Google» закрепилось крепкое второе место и в эти годы его выбирали 44-46%, оставшиеся 2-4% распределились между: Bing, Rambler и др.. Однако была замечена тенденция к росту популярности «Google» и в 2019 году 50,7% поисковых запросов в России было сделано именно через американскую компанию. Тогда как у Яндекса популярность упала до 46,6%.

Все поисковые системы имеют свои различные принципы работы, однако их можно обобщить до следующих четырех:

1. Сбор данных – с сайта собирается вся информация о контенте (как текстовом, так и визуальном);

2. Индексация – у каждого сайта есть свой инвертированный файл индекса, по которому с помощью ключевых слов поисковая система выдает результат, без данного файла в поисковой системе сайт отобразиться не сможет;
3. Поиск информации – по введенному пользователем запросу с помощью машинного обучения находятся подходящие страницы;
4. Ранжирование – основной механизм распределения позиций сайтов в выдаче поисковика, опирается на вес сайта, соответствие запросу, авторитетность и популярность сайта и т.д.

Рассмотрим алгоритм ранжирования подробнее:

Существует несколько схем распределения весов (рангов), самая популярная (алгебраическая) основана на модели TF-IDF (term frequency-inverse document frequency - частота терминов-обратная частота документа). Смысл заключается в том, максимальная частота встречающегося термина в данном документе (TF) в связке с минимальной частотой употребления во всех документах набора (IDF), дают наибольший вес документу.

$$tf(t,d) = \frac{n_t}{\sum_k n_k} \quad (1)$$

где n_t – число вхождений термина в документе,

n_k – каждое слово в документе.

$$idf(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|} \quad (2)$$

где D – число документов в коллекции,

в знаменателе - число документов из коллекции D, в которых встречается термин.

Таким образом, показатель ранга складывается из произведения данных множителей, формула представлена ниже:

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D) \quad (3)$$

Также в процессе ранжирования используется индекс цитирования — показатель, указывающий на значимость данной страницы и вычисляющийся на

основе ссылающихся страниц на данную. Изначально использовался простейший вид данного показателя – количество ссылок на данную страницу, но он был признан не объективным и заменен на индекс взвешенного цитирования (в каждой компании имеет своё название, например, «PageRank» в Google, «вИЦ» в Яндексе).

Первый алгоритм, который разработал Google для своего поискового сервиса назывался Hilltop. Появился алгоритм в 2001 году, он производил расчет и влияние PR (PageRank) на ранжирование, то есть чем больше будет ссылающихся ссылок, тем выше показатель PR.

$$PR = (1 - d) + d \sum_{i-1}^{n} \frac{PR_i}{C_i} \quad (4)$$

PR — PageRank рассматриваемой страницы,

d — коэффициент затухания (вероятность того, что пользователь, перейдет по одной из ссылок на этой странице),

PR_i — PageRank i-й страницы, ссылающейся на страницу,

C_i — общее число ссылок на i-й странице.

Огромный скачок в сфере поисковиков сделал Google, когда компания, в ноябре 2003 года выпустила новый алгоритм под названием Florida. Алгоритм стал использовать невидимые тексты и скрытые ссылки на сайте, учитывать спам ключевыми словами. Florida – родитель новой эры – SEO.

Всего через 3 месяца компания выпускает алгоритм Brandy благодаря которому поисковая система начала учитывать синонимы запросов, что существенно повысило качество поиска. Но в течении следующих 5 лет, особых изменений в поисковых алгоритмах Google не было замечено. Летом 2005 появилась персонализация поисковой выдачи, которая основывалась на истории запросов пользователя.

Летом 2007 года стремительно врывается в российский рынок поисковых систем отечественная компания Yandex. Поисковая машина Яndex-Web была анонсирована еще в 1997 году, но началом истории считается первый алгоритм поиска информации, представленный лишь в 2007. Первые алгоритмы

назывались достаточно просто «Версия 7», «Версия 8», однако в дальнейшем было решено использовать названия городов. Именно с момента прихода алгоритмов под названием «Магадан» и «Магадан 2.0» выдача запросов стала значительно чище, а продвижение сайтов и вывод их на лидирующие позиции стало в разы труднее.

В сентябре 2008 года выходит алгоритм «Находка», который начал учитывать стоп-слова (точки, тире и т.п.) в ключевых запросах; был внедрен абсолютно другой подход к машинному обучению; значительно расширен тезаурус Яндекса, поисковая система начала понимать слова, обозначающие одно то же при слитном и раздельном написании.

Арзамас/Анадырь – алгоритмы, которые в течении года получали обновление 5 раз. Главной новинкой данного алгоритма было присваивание региона, который учитывался исходя из IP хостинга, контактов на сайте или Яндекс.Каталога. Благодаря этому нововведению людям из разных городов России показывалась информация, более подходящая им. Google же применил алгоритм с учетом местоположения пользователей лишь в 2013 году. Алгоритм получил название «Venice».

В конце 2012 года Яндекс выпускает новый алгоритм под названием «Калининград», который персонализирует выдачи для поисковых запросов и подсказок. Система начала учитывать интересы пользователя и выдавать варианты, которые окажутся наиболее предпочтительны для него. Однако, данный алгоритм учитывал интересы пользователя в целом, не опираясь на сиюминутные данные и текущие запросы пользователей. Алгоритм «Дублин», написанный годом позднее исправил эту проблему и улучшил качество поиска.

Одним из самых крупных и значимых алгоритмов поисковой системы Яндекса, кардинально изменившее результаты поиска на тот момент – алгоритм «Минусинск», опускающий сайты с некачественной ссылочной массой достаточно низко. В данном алгоритме существовал так называемый порог по количеству покупных и ненадежных ссылок, их количество не разглашается компанией, известно только, что с каждой итерацией алгоритма «безопасный»

порог понижается. После запуска алгоритма биржи SEO-ссылок значительно пострадали и практически остались без работы.



Рисунок 1 - Снижение трафика после внедрения алгоритма Минусинск

В данное время наибольший интерес вызывают алгоритмы Яндекса под названиями «Палех», «Баден-Баден» и «Королев». Первый помог поисковику лучше понимать пользователей, выдавая страницы не по ключевым словам, а по запросам, что стало возможным с использованием нейронных сетей и машинного обучения, позволив модели обучаться в процессе работы. Второй алгоритм, внедренный в 2017 году, помог системе избавиться от переоптимизированных сайтов: тексты на страницах, не несущих никакой смысловой нагрузки, а лишь содержащие множество одинаковых ключей либо сильно понижались в рейтинге сайтов, либо совсем пропадали из данной поисковой системы. Таким образом пользователи попадали на страницы только с нужным интересным контентом. Алгоритм «Королев» - один из последних алгоритмов Яндекса, построенный на основе нейронных сетей и умеющий предлагать сайты не по ключам, а по смыслу. Можно сказать, что это оптимизация алгоритма «Палех», которая учитывает не только заголовки страниц, но и содержание и изображения на ней.

У основного конкурента, поисковой системы «Google», современные алгоритмы «Panda», «Penguin», «BERT» схожи по функционалу и наполнению, но реализованы по-другому. В данной работе мы не будем их подробно рассматривать, а заострим внимание на различиях в данных поисковых системах.

Одно из основных отличий систем поиска Google и Yandex – разный подход к использованию мета-тегов страниц. Google не приветствует большое количество ключей на входе, их лучше не употреблять, а словоформы – наоборот, нужны и важны в поиске. У Яндекса же, напротив, словоформы считаются за повторения и их использование нежелательно, но множественное использование ключей также не приветствуется.

Ранжирование в данных системах также реализовано по-разному. Если Google достаточно быстро реагирует на оптимизацию (до нескольких раз в день), то Яндекс меняет выдачу только во время обновлений.

Какая поисковая система подойдет конкретному пользователю – сказать невозможно, однако, как показывает статистика, приверженцы одной системы, как правило, резко критикуют вторую, поэтому выбор остается за каждым.

Список литературы:

1. Симанков В.С., Толкачев Д.М., «Методы и алгоритмы поиска информации в Интернете», ООО Издательский дом «Библио-Глобус» - 2017
2. Авторитетный сервис статистики для сайтов URL: https://www.liveinternet.ru
3. Что такое поисковые алгоритмы? URL: https://habr.com/ru/post/351930/
4. Эволюция не по Дарвину: все алгоритмы Яндекса с 2007 по 2019 год URL: https://1ps.ru/blog/dirs/2020/evolyucziya-ne-po-darvinu-vse-algoritmyi-yandeksa-s-2007-po-2019-god/
5. Список алгоритмов поисковой системы GOOGLE URL: https://evertop.pro/algorithm-google/

List of references:

1. Simankov V. S., Tolkachev D. M., "Methods and algorithms for searching information on the Internet", publishing house "Biblio-Globus" - 2017
2. Authoritative statistics service for URL sites: https://www.liveinternet.ru
3. What are search algorithms? URL: https://habr.com/ru/post/351930/
4. Evolution not according to Darwin: all Yandex algorithms from 2007 to 2019 URL: https://1ps.ru/blog/dirs/2020/evolyucziya-ne-po-darvinu-vse-algoritmyi-yandeksa-s-2007-po-2019-god/

5. List of Google URL search engine algorithms: https://evertop.pro/algorithm-google/