УДК 621.391.3

Бондарчук А. П., Залива В. В., Государственный университет телекоммуникаций

ИССЛЕДОВАНИЕ АЛГОРИТМА РАБОТЫ ПОИСКОВОЙ СИСТЕМЫ

Рассмотрены основные характеристики поисковой системы Google, ее алгоритмы, а также нововведения в формировании поисковой выдачи. Проведен анализ процесса сканирования информации, ее индексации и обработки. Исследованы изменения в обновлениях алгоритмов поиска информации начиная с 2011 года.

Ключевые слова: алгоритм поиска, Google, индексация информации, поисковые системы, обработка информации, Mobilegeddon

I. Введение

Поисковые системы являются ключевым источником распространения информации. Данные за январь 2016 года показали, что Google имеет 65% рыночной доли. Эти цифры говорят о доминирующей позиции поисковой системе на рынке. По данным на 2012 год, было подтверждено, что компания обслуживает около 3 миллиардов поисковых запросов в день. В 2016 году эта цифра значительно больше.

Так как для современного маркетинга необходимо понимания функционирования поисковых систем, в статье рассмотрены основные обновления в алгоритме Google начиная с 2011 года. Основные же факторы ранжирования держатся в секрете и компания объявляет только о самых значимых обновлениях [1].

II. Функциональные характеристики поисковая система Google

Цель работы поисковой системы – поиск введенных в форму терминов, именуемых ключевыми словами, и выдача актуальной на данный момент времени информации. С этого следует понимать, что целью каждой поисковой системы есть понимание способа поиска, как именно люди ищут нужную им информацию. Представьте, что открывая поисковую систему Google, вы попадаете в каталог, в котором отобрана лучшая информация по рейтингам, которые формирует алгоритм. Чтобы понимать с чего все начинается, необходимо рассмотреть ключевые процессы поиска.

Сканирование. Описание сканирования можно рассматривать как процесс изучения общедоступных страниц в интернете автоматизированными и систематическими методами. Когда поисковая система Google обнаруживает обновленные или новые страницы, она добавляет их в базу. Так же это можно сделать самому. С появлением сервиса Google Webmaster, у владельцев сайтов появилась возможность самим выбирать способы индексации своих веб- ресурсов. Для ускорения индексации можно загружать карту сайта в базу Google, используя Webmaster tools. Так же возможно исключение отдельных страниц сайта из индексации и запрет на сканирование. Если на сайте обнаружены ссылки на другие ресурсы, то поисковый бот так же добавляет их в свою базу и устанавливает систематические связи. Сканирование происходит регулярно, с целью установления новых связей, обновления информации и изъятия уже не существующих ссылок из базы. Масштабность этой задачи можно представить себе тем, что по данным на 2014 год в сети интернет, насчитывалось более 1 миллиарда веб- ресурсов. Но это не влияет на поведения сканирования, и боты посещают каждый веб-ресурс.

Индексация. Этот процесс идет сразу же после сканирования. Индексация – это процесс сохранения, полученной в результате сканирования информации, с последующим ее извлечением.

Особый интерес для индексации представляют ключевые слова и их расположение, мета-теги и ссылки. Так как поисковая система не видит внешний вид вашего сайта, общение с ней происходит только посредством размещения на странице вышесказанных элементов разметки.

Для хранения полученной информации используются крупнейшие центры обработки и хранения данных в Азии, Северной и Южной Америке, а также Европе. В суммарном количестве задействованы около одного миллиона серверов. Индексация служит во благо ищущего информацию человека.

Обработка. При вводе запроса в форму поиска, Google сразу же производит поиск этой информации в своей базе данных или информации, которая подходит под условия запроса. Определяя актуальность содержания, поисковая система выводит наиболее подходящие веб-ресурсы на экран пользователя.

Релевантные результаты получают высший ранг среди других остальных, менее релевантных источников информации, и выводятся на первые страницы поиска. Таким образом, обрабатывая полученную после сканирования и индексации информацию, поисковая система старается обеспечить пользователя наиболее подходящим ответом на его запрос [2].

Официальных данных по поводу методов определения релевантности нет. Google тщательно скрывает свои алгоритмы и системы. Но при всей скрытности, компания подтвердила наличие более 200 факторов определения релевантности информации и дает рекомендации веб-разработчикам, объявляя важные изменения в своих алгоритмах. Все описанные процессы обработки информации, то есть сканирования, индексирования и позиционирования, мы можем изобразить при помощи схемы:

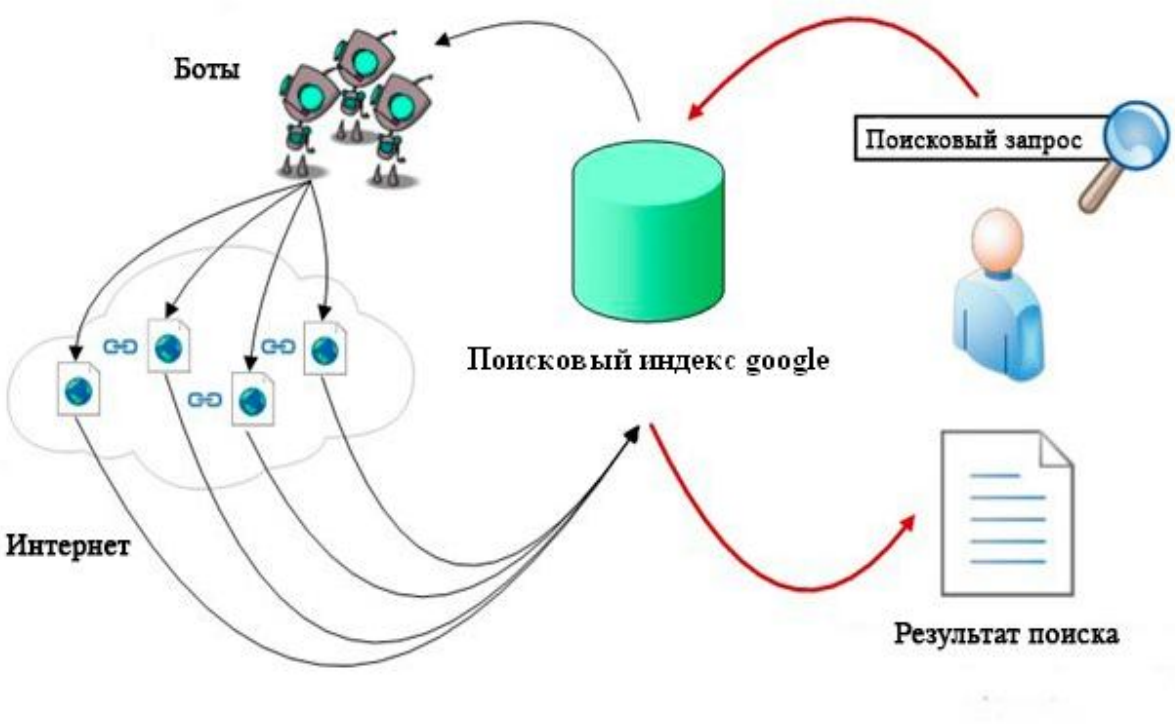

Рис. 1. Процессы обработки информации поисковой системой Google

III. Обновления алгоритмов Google с 2011 года

Алгоритм Panda. Это обновление было представлено в феврале 2011 года. В обновлении был улучшен алгоритм поиска, который нацелен на повышение качества контента. Основные стандарты качества:

- содержимое страницы должно быть больше или равно 1500 слов, иначе странице будет дан более низкий рейтинг;
- информация, которую вы предоставляете пользователям на веб-сайте, должна быть оригинальной, с большим показателем уникальности контента;
- уникальность самой темы в предоставляемом контенте;
- орфографически и грамматически правильный текст.

Обновление Page Layout (Top Heavy). Выпущенное в январе 2012 года обновление предусматривает наказание для сайтов с большим содержанием рекламных материалов в верхней части страницы. Также наказание предусматривается для сайтов с чрезмерно агрессивной рекламой, которая может отвлекать от основного содержания сайта. Такое обновление вышло в результате большого количества жалоб от пользователей, которым было трудно найти нужную информацию среди огромного количества рекламы. Таким образом Google призывает, в первую очередь, делать акцент на содержимое сайта, не мешая удобству усвоения информации.

Обновление Penguin. Алгоритм был выпущен в апреле 2012 года и направлен на борьбу с поисковым спамом. Сайты, использовавшие спам методы, были значительно понижены в рейтинге или вовсе удалены из него.

Обновление Pirate. Вышедшее на свет в августе 2012 года, это обновление понизило в рейтинге сайты, которые нарушали авторские права и интеллектуальную собственность. Правообладатели могут применять систему для сообщения о нарушениях авторских прав теми или иными веб- ресурсами.

Обновление Exact Match Domain(EMD). Выпущено в сентябре 2012 года с целью борьбы с доменами, которые созданы специально для контекстно- медийной рекламы Google AdSense.

Обновление Payday Loan. Было выпущено в июне 2013 года и направлено на уменьшение страниц, которые содержат много спама среди ключевых слов. Такая уловка использовалась веб-мастерами для продвижения сайтов среди поисковой выдачи.

Обновление Hummingbird. Обновление вышло в 2013 года и его цель состояла в том, чтобы понимать смысл поискового запроса и возвращать соответствующий результат. Это означает, что совпадение ключевых слов стало иметь меньший приоритет, чем намерение самого запроса.

Обновление алгоритма Pigeon. Обновление было выпущено в июле 2014 года и уделяло внимание проблемы геозависимого поиска. Геолокация пользователя была и есть ключевым параметром ранжирования [3, 4].

Последнее обновление Mobilegeddon. Это последнее на данный момент обновление от Google, которое влияет только на мобильный поиск. В рейтинге поисковой выдаче дается преимущество тем сайтам, которые имеют адаптивный к мобильным устройствам дизайн. Это обновление не затрагивает алгоритм поиска с Desktop компьютеров или планшетов. В отличии от прошлых алгоритмов Panda и Penguin, алгоритм Mobilegeddon работает в режиме реального времени. Компания Google создала специальный тест, в котором можно будет проверить свой сайт на совместимость с мобильными устройствами.

IV. Выводы.

Основная задача и цель Google – это двигаться в направлении лучшего обеспечения и предоставления самого высокого качества ответов на пользовательские запросы. Технические особенности всегда будут подвергаться модификациям и изменениям, в то время как общая идея и стратегия вряд ли останется тронутой. Так как поведение человека имеет свойство изменятся вместе с течением тенденций, задача Google приспосабливаться к этим тенденциям. То, что Google с этим несомненно справляется, говорит алгоритм «Mobilegeddon», который появился из-за стремительно растущего числа мобильных устройств, которые участвуют в поиске. Если вы вебмастер, то сосредоточенность на понимании базовых принципов поиска и понимании потребностей ваших клиентов, вы обязательно достигните желаемых целей.

Список использованной литературы

1. Матичин І. І., Онищенко В. В. Моделювання та аналіз трафіку в телекомунікаційних системах та мережах // Вісник Державного університету інформаційно-комунікаційних технологій. – 2013. – №. 4. – С. 20-27.

2. Chen M. Research on model of network information extraction based on improved topic-focused web crawler key technology / M. Chen, X. P. Yang // Tehnicki vjesnik/Technical Gazette. – 2016. – Т. 23. – №. 4. – PP. 1025-1035.

3. Как работает Google Поиск, основные алгоритмы обновлений [Електронний ресурс] // – Режим доступу : https://habrahabr.ru/company/ua-hosting.

4. Nasomyont T. A Study on The Relationship between Search Engine Optimization Factors and Rank on Google Search Result Page / T. Nasomyont, N. Wisitpongphan // Advanced Materials Research. – Trans Tech Publications. – 2014. – Т. 931. – P 1462-1466.

Автори статті

Бондарчук Андрій Петрович – кандидат технічних наук, доцент кафедри інженерії програмного забезпечення. Державний університет телекомунікацій, Київ. Тел.:+380 (97) 408 61 31. E-mail: 0-99@mail.ru

Залива Віталій Віталійович – студент, Державний університет телекомунікацій, Київ. Тел.: +380 (66) 480 48 57. E-mail: neapolvet@yahoo.com

Authors of the article

Bondarchuk Andrii Petrovych – candidate of sciences (technical), associate professor of software engineering departmentState University of Telecommunications, Kyiv. Tel.:+380 (97) 408 61 31. E-mail: 0-99@mail.ru

Zalyva Vitalii Vitaliyovych – student, State University of Telecommunications, Kyiv. Тел.: +380 (66) 480 48 57. E-mail: neapolvet@yahoo.com.

Дата надходження
в редакцію: 03.11.2016 р.

Рецензент:
доктор технічних наук, професор В. Є. Бондаренко
Державний університет телекомунікацій