УДК 681.3

М.А.Демихов

ХАРАКТЕРИСТИКИ АЛГОРИТМОВ ПОИСКА В СОВРЕМЕННЫХ ПОИСКОВЫХ СИСТЕМАХ

Воронежский институт высоких технологий

Рассмотрены основные свойства алгоритмов, которые применяются в современных поисковых системах. Указаны критерии оценки релевантности. Обсуждается, каким образом можно эффективно использовать зональные индексы.

Ключевые слова: поисковая система, алгоритм, критерий, релевантность, интернет.

Алгоритмы в поисковых системах являются особыми математическими формулами, на основе которых поисковыми системами решаются задачи по выдаче сайтов в результаты поиска. Анализ показывает, что поисковыми машинами ищутся сайты на основе определенных ключевых слов или фраз [1-6].

С привлечением алгоритмов поисковых систем есть возможности найти сайты, которые наибольшим образом соответствуют запросам пользователей, при этом отсекаются площадки, которые не нужны для пользователей или они используют те методы оптимизации, которые не разрешены [7-15].

В этой статье рассмотрены проблемы поисковых систем и опциональных решений.

В поисковой машине, на основе алгоритма, идет анализ контента сайтов, идет выяснение наличия в составе контента определенной ключевой фразы, принимается решение о том, насколько данный сайт может соответствовать запросу пользователя, и после этого, исходя из того, какая степень соответствия сайту присваивается та или иная позиция в выдаче - выше или ниже.

По каждой поисковой системе идет разработка своих алгоритмов. Схемы функционирования по всем алгоритмам в поисковых системах строятся на похожих принципах. В качестве примера, можно отметить, что всеми поисковиками обязательно происходит оценка контента. В алгоритмах поисковых систем есть отличия по некоторыми другим инструментам анализа.

Простейшим алгоритмом поиска является линейный. Происходит просмотр массива последовательным образом от первого до последнего элемента. При этом может быть худший случай, когда искомого слова нет в массиве, может быть определен лишь после того как просмотрен весь массив. В качестве достоинства следует указать простоту реализации. В качестве недостатка большое время, требуемое для поиска. Если мы используем оператор for, то всегда будет выполнено ровно n операций по

сравнению, причем это не зависит от того, нашли мы слово или нет. Следует, по-видимому, делать прекращение поиска в тех случаях, когда слово было найдено (если нет требований по определению всех вхождений слова).

Для равномерного распределения элементов по массиву величина среднего времени поиска обычно является пропорциональной величине n/2.

Хорошая поисковая система не пытается найти страницу, которая лучше всего соответствует запросу, она отвечает на запрос. Поэтому самые популярные поисковые системы используют сложный алгоритм, чтобы определить, какие результаты они должны получить.

В алгоритме есть "жесткие факторы", такие как число ссылок на страницу, и социальные рекомендации за ближайший год, т.е. грубо говоря, учитывается статистика перехода на страницы с похожим запросом. Это внешние факторы. Также есть факторы, относящиеся непосредственно к искомой странице, такие как методы построения элементов, которые играют большую роль в алгоритме, они относятся к внутренним факторам.

Только вместе внешние и внутренние факторы могут сформировать максимально приближенный вариант ответа на запрос.

Алгоритмы Google, Yandex и других поисковиков в точности неизвестны, но можно понять принципы их работы по внешним признакам [16].

Поисковые системы на сегодняшний день чрезвычайно развиты, однако они начинают работу с простейших логических операторов, во-первых установление истинности нахождения слова в документе, потом использование операторов OR, AND, NOT.

Лидер поисковых машин Интернета, Google занимает более 70 % мирового рынка, а значит, семь из десяти находящихся в сети людей обращаются к его странице в поисках информации в Интернете. Сейчас регистрирует ежедневно около 50 миллионов поисковых запросов и индексирует более 8 миллиардов веб-страниц.

Если говорить об истории алгоритмов в поисковой системе Google, то они начались с того, что был введен поисковый способ по индексу цитирования, который заключался в том, что ранжировались страницы и сайты в зависимости от числа и показателя авторитета ссылок (PageRank), которые ведут на них со стороны других ресурсов [17-18].

В результате сеть Интернет можно рассматривать как коллективный разум, который и определяет величину релевантности сайтов. Такую концепцию можно считать как удачным нововведением, вследствие которого Google и является весьма популярной системой.

Для ранних этапов в алгоритмах Google уделялось внимание только внутренним атрибутам страниц. Затем стали приниматься во внимание

такие факторы, как степень свежести информации и к какой географической области она принадлежит.

С 2000 г. стали использовать алгоритм Hilltop, который был предложен Кришной Бхаратом, для того, чтобы более точного проводить расчет PageRank. В следующем году полностью переписали первоначальный механизм работы систем.

После этого в Google стали делать деление по коммерческим и некоммерческим страницам. И ввели коэффициент, который позволил добавлять для ссылок, которые шли с авторитетных сайтов, большие значения весов.

Алгоритм "Флорида", который ввели в 2003 г. повлиял на распределение позиций в поиске, поскольку были исключены страницы, имеющие повторяющийся анкором и имеющие повышенное содержание ключевых слов.

С 2006 г. в системе стала использоваться разработка Ори Алона, это алгоритм "Орион", который улучшал поиск вследствие отображения тех результатов, которые были и принимал во внимание качество для тех сайтов, которые были индексированы.

С 2007 г. компанией Google стал применяться алгоритм "Austin", а с 2009 г. - анонсирован алгоритм "Caffeine.

Среди последних разработок можно отметить алгоритм «Дублин», который позволяет проводить отслеживание интересов пользователей, которые у них существуют в текущий момент.

Причем учет пользователей осуществляется буквально в течение несколько секунд.

Для критериев оценки релевантности, которые используются в поисковых алгоритмах, следует указать основные такие:

- величина возраста ресурса;
- значение уровня домена и доменной зоны;
- соответствие между ключевыми словами и содержанием сайта. Популярность для тематики;
- величина объема информации по ресурсу;
- степень близости стиля по всем страницам сайта;
- индекс цитирования;
- метод форматирования по ключевым словам в тексте;
- что содержится в метатегах;
- величина глубины сайта, число его уровней (говорят о переходах, которые мы можем делать без того, чтобы посетить одну и ту же страницу);
- существование комментариев для кода страницы.

Относительно простой метод, заключающийся в использовании зональных индексов. Веб-страница может быть разделена на несколько зон. Например - название, описание, авторы и главный блок. Мы можем разработать простую оценку для любого документа используя назначаемые для каждой зоны коэффициенты.

Этот метод начинает работать одним из первых среди остальных поисковых механизмов. Предположим, есть размеченные зоны:

Блок	Коэффициент
Заголовок "Новости"	0.4
Описание "Открытие больницы"	0.1
Тело "Сегодня, в 12-00 состоялось торжественное открытие новой больницы"	0.5

При введении запроса "больница", мы получим в первой графе 0, во второй 0.1, в третьей 0.5. В сумме 0.6 - это число будет отражать приоритет страницы. Google использует более сложный алгоритм с введением дополнительной зональности, но принцип остается таким же.

Главная трудность состоит в том, что документы имеют разную структуру. Проще всего размечать зоны в XML. Интерпретация в HTML сложнее, структура и метки более ограничены, что приводит к трудностям в анализе.

Чтобы определить контекст страницы, Google будет разделять веб-страницу на блоки. Таким образом Google будет судить, какие зоны на странице важны, а какие нет.

Один из методов, которые можно использовать - это отношение текст\код. Зоны, который содержит много ссылок / HTML кода и мало содержания, будет иметь меньший приоритет.

Использование отношения текст\код является лишь одним из методов, которые поисковая система может использовать, чтобы разделить страницу на зоны.

Преимущество метода коэффициентов зоны состоит в том, что можно рассчитать объективный балл для каждого документа. Недостатком, является то, что слишком много документов получают один и тот же балл.

Учитывая большой набор записей в базе данных и количество запросов, поиск по сходству стремится найти все данные, субъективно похожие на запрос. Чтобы решить эту проблему, необходимо рассмотреть 2 аспекта:

1. Для выполнения выборочного поиска набор данных определяется по мере сходства.
2. Эффективный метод, который будет установлен при наименьшем использовании доступа к базе данных. Сравнение идет по мере сходства.

Большинство «русскоязычных» поисковых систем индексируют и ищут тексты на многих языках — украинском, белорусском, английском, татарском и др. Отличаются же они от «всеязычных» систем, индексирующих все документы подряд, тем, что в основном индексируют ресурсы, расположенные в доменных зонах, где доминирует русский язык или другими способами ограничивают своих роботов русскоязычными сайтами.

Доля русскоязычных поисковых систем:

- Yandex (61,3 %),
- Mail.ru (8,5 %),
- Rambler (2.2 %).

В 2008 году яндекс запустил новый алгоритм "находка". Основные изменения связаны с отличиями в способе учета стоп-слов (выросло качество ранжирования по запросам со стоп-словами) и новым подходом к машинному обучению.

Заметным образом расширен тезаурус путем автоматического анализа проиндексированного корпуса текстов. Например, в нем появились сочетания слов, которые в раздельном написании означают то же самое, что и в “склеенном” виде (теперь по запросу [авто ваз] найдутся страницы и со словом “автоваз”).

Улучшено ранжирование запросов, в которых присутствуют различные союзы и предлоги. В алгоритме увеличился тезаурус, или по-другому словарь связей. Некоторая часть запросов разбавляется сайтами информационного характера, например, Википедией.

Вывод. Таким образом, при выборе алгоритмов поиска необходимо ориентироваться на критерии оценок релевантности и ранжирование запросов.

ЛИТЕРАТУРА

1. Фомина Ю.А., Преображенский Ю.П. Принципы индексации информации в поисковых системах / Вестник Воронежского института высоких технологий. 2010. № 7. С. 98-100.
2. Зяблов Е.Л., Преображенский Ю.П. Построение объектно-семантической модели системы управления / Вестник Воронежского института высоких технологий. 2008. № 3. С. 029-030.
3. Преображенский Ю.П. Разработка методов формализации задач на основе семантической модели предметной области / Вестник Воронежского института высоких технологий. 2008. № 3. С. 075-077.

4. Преображенский Ю.П. Оценка эффективности применения системы интеллектуальной поддержки принятия решений / Вестник Воронежского института высоких технологий. 2009. № 5. С. 116-119.
5. Зазулин А.В., Преображенский Ю.П. Особенности построения семантических моделей предметной области / Вестник Воронежского института высоких технологий. 2008. № 3. С. 026-028.
6. Преображенский Ю.П. Разработка методов формализации задач на основе семантической модели предметной области / Вестник Воронежского института высоких технологий. 2008. № 3. С. 075-077.
7. Zobel J., Moffat A. Inverted Files for Text Search Engines. / ACM Computing Surveys, vol. 38, № 2, 2007. - 121p.
8. Иванов М.С., Преображенский Ю.П. Разработка алгоритма отсечения деревьев / Вестник Воронежского института высоких технологий. 2008. № 3. С. 031-032.
9. Зяблов Е.Л., Преображенский Ю.П. Разработка лингвистических средств интеллектуальной поддержки на основе имитационно-семантического моделирования / Вестник Воронежского института высоких технологий. 2009. № 5. С. 024-026.
10. Чопоров О.Н., Чупеев А.Н., Брегеда С.Ю. Методы анализа значимости показателей при классификационном и прогностическом моделировании / Вестник Воронежского государственного технического университета. 2008. Т. 4. № 9. С. 92-94.
11. Завьялов Д.В. О применении информационных технологий / Современные наукоемкие технологии. 2013. № 8-1. С. 71-72.
12. Чопоров О.Н., Наумов Н.В., Куташова Л.А., Агарков А.И. Методы предварительной обработки информации при системном анализе и моделировании медицинских систем / Врач-аспирант. 2012. Т. 55. № 6.2. С. 382-390.
13. Чопоров О.Н., Агарков А.И., Куташова Л.А., Коновалова Е.Ю. Методика преобразования качественных характеристик в численные оценки при обработке результатов медико-социального исследования / Вестник Воронежского института высоких технологий. 2012. № 9. С. 96-98.
14. Малышев В. А. Основные проблемы научных исследований при разработке и совершенствовании технических систем / Вестник Воронежского института высоких технологий. 2015. № 14. С. 8-11.
15. Питолин М. В., Мачтаков С. Г. Постановка задачи классификации ассоциативного поиска объектов в базе данных / Вестник Воронежского института высоких технологий. 2015. № 14. С. 37-39.
16. Ахо Альфред, Хопкрофт В., Джон, Ульман, Джеффри Д. Структуры данных и алгоритмы. - Издательский дом "Вильямс", 2000. - 384 с.

17. Бойцов Л.М. Использование хеширования по сигнатуре для поиска по сходству. Прикладная математика и информатика. / ВМиК МГУ, 2011, № 8, с. 135-154.
18. Скородумов В. А., Соколовский В.В. Обзор задач и методов смысловой обработки электронных данных/ Интернет в библиотеке : Ежегод. межведомств. сб. науч. тр. / Гос. публ. науч.-техн. б-ка России ; [Редкол.: В. А. Скородумов и др.]. - М. : ГПНТБ России, 2003. - С. 79-85.

M.A.Demikhov

THE CHARACTERISTICS OF SEARCH ALGORITHMS IN MODERN SEARCH ENGINES

Voronezh institute of high technologies

The main properties of the algorithms used in modern search engines is considered. The criteria for assessing relevance are pointed out. It was discussed how we can effectively use the zonal indices. .

Keywords: search engine, algorithm, criterion, relevance, Internet.