УДК

ГРНТИ

СРАВНИТЕЛЬНЫЙ АНАЛИЗ НЕКОТОРЫХ ПОИСКОВЫХ СИСТЕМ

Зиен Роман Маратович

Студент

Санкт-Петербургский горный университет

Россия, г. Санкт-Петербург

Аннотация

Была сравнена эффективность поиска информации в некоторых популярных поисковых системах (а именно, Google, Yahoo, AlltheWeb, Gigablast, Zworks и AltaVista и Bing / MSN) в ответ на список из десяти запросов разной сложности. Эти запросы выполнялись в каждой поисковой системе, и точность и время отклика полученных результатов были записаны. Первые десять документов при каждом поиске результаты были оценены как "релевантные" или "нерелевантные" для оценки поисковой системы точность. Чтобы оценить время отклика, были рассчитаны нормализованные коэффициенты вспоминания при различных пороговых значениях и давались баллы за каждый запрос и поисковую систему. Это исследование показывает, что Google – имеет лучший поисковый движок запросов с точки зрения как средней точности (70%), так и среднего времени отклика (2 с). Gigablast и AlltheWeb показал худшие результаты в этом исследовании.

Ключевые слова: каталог; индекс; точность; время отклика; поисковый движок; паук

COMPARATIVE ANALYSIS OF SOME SEARCH ENGINES

Zien Roman Maratovich

Student

Saint-Petersburg Mining University

Russia, St. Petersburg

ABSTRACT

We compared the information retrieval performances of some popular search engines (namely, Google, Yahoo, AlltheWeb, Gigablast, Zworks and AltaVista and Bing/MSN) in response to a list often queries, varying in complexity. These queries were run on each search engine and the precision and response time of the retrieved results were recorded. The first ten documents on each retrieval output were evaluated as being ‘relevant’ or ‘non-relevant’ for evaluation of the search engine’s precision. To evaluate response time, normalised recall ratios were calculated at various cut-off points for each query and search engine. This study shows that Google appears to be the best search engine in terms of both average precision (70%) and average response time (2 s). Gigablast and AlltheWeb performed the worst overall in this study.

Keywords: catalogue; index; precision; response time; search engine; spider.

Вступление

Поиск во всемирной паутине стал частью нашей повседневной жизни, поскольку Интернет теперь является необходимым инструментом для сбора информации и, несомненно, обеспечивает удобство поиска информации, потому что он может объединять информацию с разных веб-сайтов. Акередолу официально описал Сеть как «глобальную инициативу по поиску информации в гипермедиа, цель которой - предоставить универсальный доступ к большим количествам документов». Проще говоря, Интернет - это компьютерная сеть на базе Интернета, которая позволяет пользователи на одном компьютере для доступа к информации, хранящейся на другом, через всемирную сеть. Конечная цель разработки и публикации веб-страницы - поделиться информацией. Тем не менее, большое количество веб-страниц,

ежедневно добавляемых в Интернет, превратило Интернет в море всех видов данных и информации, которая затрудняет поиск информации. Количество информации в Интернете, а также количество хостов и доменных имен, зарегистрированных во всем мире, быстро растут. В настоящее время существует более 1 триллиона веб-страниц, и, по оценкам, количество веб-страниц будет продолжают расти до бесконечности. Ежедневно в Интернет добавляется несколько миллиардов веб-страниц. Эта новая информация должна быть доступна для всех, чтобы веб-страница могла достичь намеченной цели. Чтобы преодолеть эти проблемы поиска, более 20 компаний и учреждений разработали инструменты поиска, такие как Yahoo, AltaVista, Google и Lycos.

Инструменты поиска можно разделить на два основных типа: каталоги и поисковые системы. Главное отличие между каталогами и поисковыми системами заключается в том, что каталог создается людьми, в то время как поисковые системы – это база данных, создается программным обеспечением, известным как пауки или роботы. Поиск вместо просмотра – это главная особенность поисковых систем. Преимущество поисковых систем перед каталогами в том, что они очень дают исчерпывающий результат, часто включающий в список результатов тысячи сайтов. Таким образом, недостатком является то, что пользователю нужно просмотреть тысячи нерелевантных сайтов, чтобы найти то, что он ищет, потому что, хотя поисковые системы пытаются перечислить сайты в порядке релевантности, эта релевантность определяется математическими формулами, которые далеки от совершенства. Поисковые системы особенно полезны при поиске определенной темы, которая может отсутствовать в каталоге. Поисковые системы - это инструменты для поиска информации и каталогов - это коллекции проверенных людьми веб-сайтов, сгруппированных по тематическим категориям. Следовательно, каталоги влияют на поисковые системы, отсюда и взаимозаменяемость. Хотя у них разные стратегии поиска, как поисковые системы, так и каталоги имеют схожие интерфейсы и широко

известны в качестве поисковых систем; поэтому в этой статье оба типа называются поисковыми системами.

Поисковые системы входят в список наиболее посещаемых сайтов. Поисковые системы создают и поддерживают индекс слов в документах в Интернете. Они возвращают пользователю ранжированный список релевантных документов как результаты поиска. Некоторые результаты могут быть ценными для пользователя, тогда как большинство обычно не имеет значения.

По мере того как Интернет продолжает расти, большинство поисковых систем сталкиваются с серьезными проблемами; Интернет растет намного быстрее, чем может индексировать любая существующая технология, и поэтому результаты могут стать устаревший. Многие веб-страницы часто обновляются, что заставляет поисковые системы посещать их периодически. Многие динамически генерируемые сайты также не индексируются поисковыми системами. Самые большие поисковые системы (то есть с самыми большими индексами) проделали впечатляющую работу по расширению своего охвата, но технология должна была значительно масштабироваться, чтобы идти в ногу с ростом Интернета.

В 1994 г.World Wide Web Worm, одна из первых поисковых машин в Интернете, имела индекс в 110 000 веб-страниц и документов, доступных в Интернете. По состоянию на 2006 год Google проиндексировал 25 миллиардов веб-страниц; в настоящее время, в 2020 г.,Google индексирует 29,2 миллиарда веб-страниц каждый день. В то время как количество веб-страниц увеличивается, количество запросов, которые поисковым системам требуется обрабатывать также невероятно растет. В настоящее время Google получает в среднем 400 миллионов запросов в день, а по состоянию на июль 2020 года AltaVista заявила, что обрабатывала около 13 миллионов запросов в день. Учитывая проблемы, с которыми сталкивается большинство поисковых систем, потребность в улучшенных поисковых системах и быстро находить нужную информацию для удовлетворения различных потребностей

пользователей Интернета становится все более актуальной. Правильный выбор поисковой системы помогает уменьшить трудности, возникающие при поиске информация из сети. Столкнувшись с таким количеством поисковых систем, пользователи могут легко запутаться.

Пользователи обычно возвращаются к одной или двум поисковым системам с которые им удобно работать. Однако какая поисковая система действительно удовлетворяет потребности пользователя и какая из них лучше? Отвечая на эти вопросы, пользователь должен уточнить свои потребности и какие функции он предпочитает, например, объем полученной информации, скорость получения результатов поиска или релевантность результатов поиска.

В этом исследовании использовался эмпирический подход для оценки точности и скорости поиска информации в некоторых выбранных поисковые системы. Результаты позволят пользователям лучше понимать возможности поисковой системы, делать выводы о различных поисковых системах и открыть для себя возможности для дальнейшего поиска информации.

Как работают поисковые системы?

Существуют разные способы организации веб-контента, но каждая поисковая система имеет те же основные части, что и сканер или паук, индекс или каталог, а также интерфейс или запрос-модуль1. Пользователи вводят поисковый запрос через заранее заданный модуль запросов,

[1] Значение терминов
Поисковая система: программа, которая ищет документы по указанным ключевым словам и возвращает список документов, в которых были найдены ключевые слова, ранжированные в порядке актуальности. Это позволяет запрашивать контент, отвечающий определенным критериям (обычно теме, содержащее заданное слово или фразу) и получает список ссылок, которые соответствуют этим критериям.
Каталог: ручной каталог сайтов в Интернете. Люди создают категории и назначают сайту место в структурированном индексе. Пример типового каталога - это Yahoo, который проверяет всю соответствующую информацию и направляет эту информацию по адресу. Yahoo также показывает сайты так, чтобы в каждой категории появляется первым в списке. Эта функция поиска может помочь людям быстро найти нужную информацию по более общим темам.
Сканер / Паук: посещает веб-страницы по ссылкам, обновляет страницы и добавляет новые страницы, когда он их встречает.
Индекс / Каталог: место, где хранятся данные, собранные пауком, т. е. Содержит копию каждой веб-страницы, которую находит паук.
Запрос: ключевое слово или вопрос, введенный пользователем, в поисковую систему.
Время ответа: период между выдачей поискового запроса и отображением первых результатов поиска.
Точность: соответствие результата поиска поисковому запросу.

специфичный для каждой поисковой системы. Обычно поисковая система работает, посылая паука, чтобы получить как можно больше документов по возможности. Затем другая программа вызывает индексатор, который читает эти документы и создает индекс на основе слова, содержащихся в каждом документе. Каждая поисковая система также использует собственный алгоритм для создания индексов, которые в идеале позволяют возвращать только значимые результаты для каждого запроса. Как правило, поисковая система начинает с набора предопределенных веб -адреса и скачивает их. Для каждой страницы она извлекает единый указатель ресурса или URL-адрес, чтобы следить за ними позже указанным способом. Затем она индексирует все слова и фразы, а также взаимное расположение слов друг к другу. Затем пользователь может выполнить поиск по этому индексу по полученным результатам на наличие определенного слова, фразы или сочетания слов в веб-документе.

МЕТОДЫ

Десять поисковых запросов использовались для тестирования семи различных запросов поисковых систем; как точность, так и время отклика поиска. Полученные результаты затем сравнивались с результатами поисковых систем.

Подбор поисковых систем

Поисковые системы, выбранные для сравнения в этом исследовании, были Yahoo, Google, Gigablast, AlltheWeb, Zworks, AltaVista и Bing / MSN.B процессе выбора поисковых систем, которые будут оценены, внимание было уделено включению разнообразного диапазона поисковых системы, чтобы полученные результаты могли служить основой для оценки алгоритма поиска, используемого различными поисковыми движками. Некоторые из выбранных поисковых систем не самые популярные. Таким образом, результаты исследования будут информировать пользователей об их различных возможностях и тем самым потенциально увеличат использование более эффективных поисковых движков. Единые поисковые системы в

Интернете такие как CUSI (Configurable Unified Search Index) не считались, потому что они только компилируют существующую веб-информацию и не предоставляют ничего нового.

Тестовые запросы

Десять поисковых запросов были разработаны для использования во всех запросах поисковых движков. Эти запросы были разработаны для тестирования различных функций поисковой системы, а также они представляют разные уровни сложности поиска. Запросы также были предназначены для использования в сфере информационных технологий с целью ознакомления, чтобы следователи могли оценить результаты поиска на предмет релевантности. Десять запросов были разделены на четыре группы следующим образом:

Таблица 1A

Время отклика поисковых систем, измеряемое в секундах, на четыре выбранных запроса различной сложности в непиковые часы

Запрос	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN	Средн ее	с. о.
1	6	2	3	9	5	6	2	5	3
6	10	2	2	10	5	5	4	5	3
8	8	2	7	6	5	10	2	6	3
10	7	3	8	9	4	6	3	6	2
Средн ее	8	2	5	9	5	7	3	н/д	н/ д
с.о.	2	1	3	2	1	2	1	н/д	н/ д

с.о.* - стандартное отклонение

Таблица 1B

Время отклика поисковых систем, измеряемое в секундах, на четыре выбранных запроса различной сложности в часы пик.

Запрос	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN	Средняя	с.о.
1	38	12	18	25	9	15	17	19	8
6	12	9	25	38	17	27	15	21	8
8	25	23	18	24	13	21	15	19	4
10	23	18	17	32	24	25	14	21	4
Средняя	25	16	20	30	16	22	15	н/д	н/д
с.о.	9	5	3	6	6	5	1	н/д	н/д

Таблица 2

Оценка точности * для каждого запроса, выполненного в каждой поисковой системе

Запрос	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN	Средняя
1	0,8	0,7	0,7	0,6	0,6	0,7	0,5	0,7
2	0,9	0,6	0,6	0,6	0,3	0,6	0,7	0,6
3	0,7	0,8	0,3	0,8	0,7	0,6	0,8	0,7
4	0,3	0,8	0,5	0,5	0,6	0,5	0,5	0,5
5	0,6	0,8	0,6	0,7	0,5	0,5	0,6	0,6
6	0,7	0,8	0,4	0,5	0,5	0,7	0,6	0,6
7	0,7	0,9	0,3	0,6	0,5	0,6	0,5	0,6
8	0,5	0,6	0,5	0,5	0,3	0,4	0,6	0,5
9	0,5	0,5	0,2	0,3	0,4	0,3	0,3	0,4
10	0,4	0,4	0,2	0,3	0,2	0,1	0,3	0,3
Средняя	0,6	0,7	0,4	0,5	0,5	0,5	0,5	н/д
с.о.	0,6	0,5	0,6	0,6	0,5	0,5	0,5	н/д

Оценка точности* - Точность рассчитывалась как значение от 0 до 1, где 1 соответствует десяти из десяти результатов поиска, соответствующих релевантным.

Таблица 3

Ранжирование поисковых систем по времени отклика, оценке точности и общей производительности

Критерии	Поисковый движок						
	Yahoo	Google	Gigablast	AlltheWeb	Zworks	AltaVista	Bing/MSN
Время отклика	6	1	3	7	3	5	2
Точность	2	1	7	3	3	3	3
Средний ранг	4	1	5	5	3	4	2,5
Рейтинг*	4-ый	1-ый	7-ой	7-ой	3-ий	4-ый	2-ой

* - Общий рейтинг производительности на основе среднего рейтинга

А. Короткие запросы:

- Что такое интеллектуальный анализ данных? (Запрос 1)
- Веб-браузеры (запрос 2)
- Нейронная сеть (запрос 3)
- Эволюция микропроцессора (Запрос 4)
- Поиск по ключевым словам (запрос 5)

Б. Логические запросы (И / ИЛИ):

- Поиск И сортировка (Запрос 6)
- Алгоритм кластеризации ИЛИ кластеризации (запрос 7)

С. Запросы на естественном языке:

- Поиск в Интернете с использованием естественного языка (запрос 8)

- Как добиться наилучших результатов поиска в Интернете? (Запрос 9)

D. Длинный запрос:

- Я нашел классную веб-страницу, но потерял ее. Как мне его вернуть? (Запрос 10)

Для каждого запроса оценивались только первые десять результатов поиска. Для большинства пользователей первые десять полученных результатов являются наиболее важными, т.е. почти все пользователи надеются, что первые десять поисковых результатов предоставят то, что они ищут, а если это не так в этом случае они разочаровываются и обычно пробуют поискать еще раз другой движок.

Учитывая, что все выбранные поисковые системы отображают результаты по убыванию релевантности, считается, что это методология не оказала критического влияния на достоверность результатов.

Тестовая среда

Microsoft Internet Explorer был выбран в качестве веб-браузера для исследования, потому что он совместимо со всеми поисковыми системами и является наиболее широко используемым браузером на местном уровне. Два компьютера с разными конфигурациями, но с одинаковыми параметрами: компьютер Acer с процессором Intel Celeron M 440, жесткий диск 80 ГБ (частота 1,86 ГГц) и 52 МБ Память DDR2 и компьютер Hewlett Packard (2,10 МГц) с процессором AMD Semipro SI-42, жесткий диск 140 ГБ и 1 ГБ оперативной памяти. Один компьютер использовался для всего эксперимента, который был повторен для достоверности на втором компьютере, т.е. каждый запрос выполнялся дважды. Показанные результаты получены с компьютера Hewlett Packard. Полученные результаты из повторного определения не представлены, потому что они были сопоставимы и не влияли на результаты исследования.

В идеале каждый запрос должен выполняться во всех поисковых системах, но в то же время, так что, если добавляется соответствующая страница, ни одна из них не должна иметь преимущество в том, что можно

проиндексировать новую страницу по сравнению с другими. Для данного исследования это было практически невозможно, поэтому каждый запрос просматривался во всех поисковых системах в пределах тридцати минут друг от друга в один и тот же день. Эти поисковые системы возвращает ошибку «404» (т.е. путь не найден) или «603» (т.е. сервер не отвечает) были отмечены для возврата. Повторные посещения производились в разное время дня, чтобы проверить сайт на вероятность того, что он может регулярно поддерживаться.

Время отклика

Время отклика рассчитывалось как период между вводом поискового запроса и получением первых результатов поиска и были измерены секундомером. Был выбран по одному запросу из каждой группы для оценки времени отклика. Были выбраны следующие запросы: Запрос1 (группа A), запрос 6 (группа B), запрос 8 (группа C) и запрос10 (Группа D). Среднее время ответа для каждой поисковой системы и для каждого выбранного запроса были затем рассчитаны.

Точность

В этом исследовании точность определялась как релевантность поиска результата поискового запроса и определялась отдельно обоими исследованиями за первые десять результатов поиска. Было проверено содержание каждого полученного результата, чтобы определить, удовлетворяет ли ожиданиям результат, но не было попыток прочитать полный текст Веб-документа, переходя по ссылкам, предоставленным из-за соображения времени и переменная надежность ссылок. Оценка точности рассчитывалась на основе количества результатов в пределах первых десяти считающихся релевантными (т.е. оценка 1 указывает, что все десять результатов поиска были релевантными, а оценка 0,5 означает, что только пять из первых десяти результатов были релевантными).Чтобы оценить общую производительность каждой поисковой системы было оценено, а не только вычислен средний балл точности для каждого запроса, но также

вычислен средний балл точности, на основе всех десяти запросов для каждой поисковой системы.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Время отклика

Среднее время ответа для всех поисковых систем было в пределах диапазона от 2 до 9 с в непиковые часы. В часы пик среднее время отклика увеличилось до 15 с и достигло 30 с. Индивидуальное и среднее время отклика для каждой поисковой системы и для каждого запроса в непиковые и пиковые часы показаны в Таблицах 1A и 1B соответственно.

Точность

Оценка точности для каждого запроса в каждой поисковой системе сведены в таблицу 2. Средние оценки точности для каждого движка поиска составляла от 0,4 до 0,7. Хотя рейтинг оценки точности варьировались между поисковыми системами в зависимости от запроса, Google получил наивысший средний балл точности 0,7, в то время как Yahoo получил второй по величине показатель точности 0,6. Gigablast получил самую низкую оценку точности 0,4. Наивысшая оценка точности для запроса 10 (т. е. Длинного запроса) была 0.4 (Google и Yahoo), что означает, что поисковым системам было сложнее обрабатывать длинные запросы по сравнению с более короткими запросами. Этот результат означает, что пользователи, желающие получить релевантные результаты поиска должны быть как можно точнее задавать поисковые запросы, содержащие только самые важные термины.

Общая производительность

В таблице 3 показаны семь поисковых систем, ранжированных по их времени отклика (от самого короткого до самого длинного) и оценки точности(от высшего к низшему), с рейтингом 1, обозначающим лучший исполнитель. Среднее значение обоих рейтингов указывает на общую производительность каждой поисковой системы.

ЗАКЛЮЧЕНИЕ

Как по времени отклика, так и по точности Google оказался лучшим исполнителем из всех оцененных поисковых систем. Следовательно, это рекомендуемая поисковая система. MSN / Bing - второе место, тоже рекомендуется. Gigablast и AlltheWeb были худшими исполнителями в этом исследовании.

Список использованной литературы

1. Waheed IB, Coop L, Kogan M. Integrated pest management (IPM) and Internet-based information delivery systems. Neotrop Entomol. 2003;32(3):373–383.
2. Akeredolu GF. Internet search tools: A comparative study and analysis. MSc thesis, Ibadan, University of Ibadan, 2005.
3. Lancaster FW, Fayen EG. Information retrieval on-line. Los Angeles: Melville Publishing Co; 1973.27. L e i g h t o n HV. Performance of four World Wide Web (WWW) index services: InfoSeek, Lycos, WebCrawler, and WWW Worm [homepage on the Internet]. c1995 [cited 2006 Jan 7]. Available from: http://www.curtin.edu.au/curtin/library
4. Alpert J, Hajaj N. We know web was big. Web search infrastructure team. [homepage on the Internet]. c2008 [cited 2008 July 25]. Available from: http://googleblog.blogspot. com/2008/07/we-know-web-was-big.html
5. Liu H, Weber RR. Web crawler [homepage on the Internet]. c2010 [cited 2010 June 7]. Available from: http://en.wikipedia. org/wiki/Web_crawler
6. Silverstein C, Pederson JO. Almost-constant time clustering of arbitrary corpus subsets. Paper presented at: SIGIR ’97. Proceedings of the 20th International ACM/SIGIR Conference on Research and Development in Information Retrieval; 1997 July 27–31; Philadelphia, PA. New York: ACM Press; 1997. p. 60–66.

7. Tillman HN. Evaluating quality on the Net [homepage on the Internet]. c2003 [cited 2003 March 28]. Available from: http://www.hopetillman.com/findqual.html

8. Smith N. Google images. Product manager Google images [homepage on the Internet]. c2010 [cited 2010 July 25]. Available from: http://googleblog.blogspot.com/search/label/search

9. Google search [homepage on the Internet]. c2010 [2010 July 20]. Available from: http://en.wikipedia.org/wiki/google_search

10. AltaVista [homepage on the Internet]. c2010 [2010 July 20]. Available from: http://en.wikipedia.org/wiki/Altavista

11. Westera G. Comparison of search engine user interface capabilities [homepage on the Internet]. c2002 [cited 2006 Jan 7]. Available from: http//www.curtin.edu.au/curtin/ library/staffpage/gwpersonal

12. Leonard AJ. Where to find anything on the Net [homepage on the Internet]. c1996. [2006 Jan 7]. Available from: http://www.emeraldinsight.com/journals.htm?articlerd

13. Liu K, Yu C, Meng W. Discovering the representative of a search engine. Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM ’02); 2002 Nov 4–9; Mclean, Virginia. New York: ACM Press; 2002. p. 558–565.

14. Singh R. Performance of World Wide Web search engines: A comparative study. Vision 2. Knol. [homepage on the Internet]. c2008 [cited 2008 July 27]. Available from: http:// knol.google.com/k/performance-of-world-wide-websearch-engines-a-comparative-study#

15. Harman D. Overview of The Second Text Retrieval Conference (TREC-2). Inf Process Manag. 1995;31(3):271–289.

16. Heting C. Information representation and retrieval in the digital age (ASIST Monograph Series). New Jersey: Information Today; 2003.

17. Courtois M, Berry MN. Results ranking in Web search engines. New York: Lukas; 1999.

18. Прохорова А.М. Регистрация сервера в поисковых системах как часть инструмента поисковой оптимизации сайта / А.М. Прохорова // Наука, техника и образование. 2016. №8 (26).

19. Ловина В.В. Исследование средств оптимизации системы продвижения сайтов /В.В. Ловина // Проблемы Науки. 2016. №18 (60).

20. Прокимнов Н.Н. Технологии использования информационных ресурсов Интернета / Н.Н. Прокимнов // Прикладная информатика. 2008. №5.