Методы сравнительного анализа современных поисковых систем и определения объема Рунета

© Сегалович И.В, Зеленков Ю.Г., Нагорнов Д.О.

Яндекс, Москва
{iseg, yuryz, den}@yandex-team.ru

Аннотация

В представленной работе рассмотрены автоматические методы сравнения количественных и качественных характеристик русскоязычной части индексов современных поисковых систем и определения объема Рунета.

Актуальность исследования определяется чрезвычайно быстрыми темпами роста русскоязычной части веба. Так, по результатам, полученным авторами в феврале 2004 года (начало систематических наблюдений), объем Рунета составлял примерно 400 млн. документов, тогда как к концу 2005 года эта цифра была уже порядка 2500 млн., т.е. увеличилась более, чем в 6 (!) раз за 2 года. Судя по последним измерениям, такой темп может сохраниться и в ближайшее время.

Новизна работы состоит в том, что в ней впервые выполнена оценка объема русскоязычного сегмента Сети, и систематически рассмотрены не только количественные, но и качественные параметры («чистота» и «свежесть») архивов наиболее крупных поисковых систем, таких как Яндекс, Рамблер, Google и Yahoo! Кроме того, разработанные авторами методы, в отличие от зарубежных аналогов, позволяют производить измерения веба на регулярной (ежедневной) основе, что особенно важно, учитывая высокую динамику изменения состояния современного интернета.

1 Введение и обзор ключевых работ по исследуемой тематике

В работе Д. Брэйка [1] убедительно показано, что ни одна из поисковых машин никогда не будет в состоянии собрать и поддерживать в актуальном состоянии полную информацию о составе и структуре всего интернета. Поэтому весьма актуальными представляются статистические исследования, посвященные разработке методов наилучшего сэмплирования веба и приближенной оценке относительных размеров баз данных современных поисковых систем, а также определения объема статического общедоступного (не скрытого за поисковыми формами запросов к онлайновым базам данных и не требующего привилегированного доступа) интернета.

Одни из первых наиболее значимых результатов в этой области были получены в работах К. Бхарата и А. Бродера из Центра системных исследований DEC [2, 3]. Основная идея их подхода состояла в том, чтобы взяв за основу большую выборку (порядка 300 тыс.) документов из каталога Yahoo!, составить по ней представительный частотный словарь (более 400 тыс. слов) максимально близко отражающий как лексический состав, так и реальные количественные пропорции между словами в интернете. Авторы исходили из предположения, что все поисковые системы должны использовать Yahoo! как один из основных информационных ресурсов при создании алгоритмов обхода веба. К сожалению, правдоподобность такого предположения не проверялась.

На основе данного словаря были созданы массивы запросов - четырехсловные дизъюнктивные и двухсловные конъюнктивные, с помощью которых генерировалась большая серия (порядка 20 тыс.) случайных запросов к некоторой поисковой системе. Из результатов поиска по каждому запросу отбирались первые 100 URL'ов, из которых случайным образом выбирался один URL. Недостатком такого подхода является зависимость полученных результатов от конкретных алгоритмов ранжирования выдач, используемых различными поисковиками.

Выбранный URL проверялся на вхождение в индекс всех поисковиков с помощью специально построенного «строгого запроса». Для этого документ с данным URL'ом загружался из веба и из него выбирался некоторый фиксированный набор слов с максимальным IDF.

После этого вычислялась доля выдачи каждой поисковой системы по отношению к системе, тестируемой в данный момент, т.е. определялись

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

попарные пересечения выдач выбранной машины со всеми другими. Затем в качестве тестируемой выбиралась другая машина, и весь процесс повторялся снова. По окончании расчетов для каждого поисковика были известны его доли в выдаче других поисковиков.

Важным результатом было установление того факта, что доли каждой машины в выдаче других систем примерно одинаковы. Это позволило сделать вывод о независимости алгоритмов построения индексов различными поисковиками, что, в свою очередь, дало возможность легко определить объем статического веба, предположив, что доля определенной поисковой системы по отношению к всему вебу равна ее доли (значение которой известно) по отношению к другим машинам. Тогда, зная абсолютный размер архива какой-либо одной из систем (из общедоступных источников или используя привилегированный доступ), сразу же можно вычислить размеры остальных машин и объем всего интернета.

Основными достоинствами работ К. Бхарата и А. Бродера [2, 3] являются очень четкая формулировка проблемы и ее нацеленность на получение конкретных результатов, воспроизводимость проведенных экспериментов любыми заинтересованными исследователями и практически полная их (экспериментов) автоматизация на всех этапах выполнения расчетов.

К сожалению, в одной из этих работ ([3]) допущен ряд неточностей в отношении оценки аналогичных результатов С. Лоренца и С. Джайлса из исследовательского института NEC, чья работа [4] рассматривается ниже.

Во-первых, больший объем статистической выборки не может быть единственным аргументом в пользу ее (выборки) большей точности (достаточно вспомнить классический пример с предвоенными выборами президента США Ф. Рузвельта). С. Лоренц и С. Джайлс [4] использовали строгий математический метод доверительных интервалов для обоснования размера своих выборок и поэтому их цифры выглядят вполне убедительно.

Во-вторых, авторы [4], вопреки утверждениям К. Бхарата и А. Бродера, для оценки объемов баз поисковых машин использовали не нормализованные URL'ы, а тексты документов, загруженных из веба.

И, наконец, С. Лоренц и С. Джайлс не включали повторно в расчеты страницы, имеющие более одного URL'а, хотя именно это и утверждают авторы [3]. Все эти положения в явном виде содержатся в [4] и совершенно непонятно как К. Бхарат и А. Бродер их не заметили. Поэтому оценки, содержащиеся в [4], вполне успешно могут конкурировать с результатами исследований [2, 3], да и расхождения в числах не так уж и велики.

Как уже упоминалось выше, другой важной работой по данной тематике является исследование С. Лоренца и С. Джайлса из NEC [4]. В экспериментах авторов по сравнению относительных размеров индексов поисковых систем использовался специально подготовленный научными сотрудниками института массив из 575 «хороших» запросов. Таким образом, здесь неявно предполагалось примерное равенство между отношением объемов «хороших» документов в базах сравниваемых поисковиков и «обычных», используемых ежедневно рядовыми пользователями. На практике такой уклон приводит к некоторому уменьшению нижней границы приводимых оценок (см. [2]).

На результаты, возвращаемые поисковыми машинами, накладывались определенные ограничения. Во-первых, списки URL'ов документов, возвращаемых в качестве результатов поиска всеми системами, должны иметь объем, не превышающий 600 элементов. Это условие необходимо для того, чтобы иметь, с одной стороны, возможность реального просмотра этих списков целиком, а с другой - избежать последствий неодинакового ранжирования выдач различными поисковиками.

Во-вторых, документы из списка включались в расчет только при условии, что они содержали точные совпадения со словоформами запроса, а не с их словоизменительными вариантами или, например, синонимами. Данное ограничение позволяет исключить отличия в алгоритмах индексирования у разных поисковых машин, связанные с применением морфологического анализа, использованием тезауруса, списков стоп-слов и т.п. Из полученных результатов также исключались документы-дубликаты, включая одинаковые страницы с разными URL'ами.

Для определения объема веба использовались 302 запроса из 575, которые возвращали не менее 50 для каждой из исследуемых систем.

Проверка на независимость индексов поисковых машин не проводилась, а заранее предполагалась.

Работа С. Лоренца и С. Джайлса [4] является качественным научным исследованием, содержащим много интересных идей, эвристик и алгоритмов. К недостаткам следует отнести, в первую очередь, довольно высокий процент ручного труда на различных этапах экспериментов, что при современных объемах интернета и архивов поисковых машин делает практически невозможным повторение проведенных экспериментов в чистом виде.

Одной из самых последних работ в этой области является исследование Z. Bar-Yossef и M. Gurevich [6] из университета Technion в Хайфе (Израиль). Главная задача, которую они поставили перед собой – разработка новых методов сэмплирования индексов поисковых машин, свободных от основных недостатков, присущих подходу К. Бхарата и А. Бродера [2]: недостаточная обоснованность значений частот терминов запросов и смещенность выборок в сторону длинных и высокоранжированных докуметов.

Для решения этой задачи авторы [6] предложили два метода: словарный метод формулировки запросов к поисковикам с учетом их (запросов) объема и метод случайных блужданий по виртуальному графу найденных документов. Основное отличие от подхода в [2] заключается в приписывании каждому анализируемому документу вероятности (веса) быть выбранным, что позволяет формировать однородные репрезентативные выборки.

Приписывание весов документам производится на основе оригинального моделирования заранее заданного случайного распределения по методу Монте-Карло, которое сначала выполняется для запросов, а потом для найденных документов.

Метод случайных блужданий заключается в поизвольном (в соответствии с некоторым заданным распределением) выборе текста запроса из некоторого начального документа, последующем обращении с этим запросом к поисковой машине, случайном выборе одного из найденных документов, произвольном выборе нового текста запроса из этого документа и т.д. Важно отметить, что данный метод вообще не требует никакого словаря для своей реализации.

Еще одним отличием данной работы является использование в качестве текста запросов не отдельных слов или их случайных сочетаний как в [2], а довольно длинных – 3-х и 5-тисловных связных фраз. По мнению авторов это дает наиболее точные результаты.

Основным недостатком этого исследования является его чрезмерная алгоритмическая сложность. Достаточно сказать, что для обоснования своих результатов авторам понадобилось написать свою (!) поисковую систему. Другими недостатками предложенных методов являются большие временные затраты на проведение экспериментов, их высокая трудоемкость и ресурсоемкость. Так, например, для получения результатов по методу случайных блужданий потребовалось выполнение 5 млн. запросов к каждому (!) исследуемому поисковику, в для словарного метода использовлись 6 двухпроцессорных машин с 2 Гб оперативной памяти каждая, массив документов в сотни гигабайт и набор из 1 млн. запросов к поисковым системам.

Кроме того, как показывает опыт разработки поисковой системы Яндекс, при использовании в качестве текста запросов многословных фраз практически невозможно добиться согласованного поведения различных поисковых систем и, поэтому, достоверность полученных данных не может быть слишком высокой.

Подводя общий итог, можно сказать, что основным недостатком всех проанализированных работ [2, 3, 4, 6] является их преимущественно «количественная» направленность, связанная с определением только размеров индекса поисковых систем и объема интернета. Важные «качественные» параметры веба и документальных архивов поисковиков, такие как доля полудубликатов (показатель «чистоты») или процент устаревших ссылок (показатель «свежести») остались практически совсем не рассмотренными. Кроме того, все предложенные методы требуют довольно значительных временных затрат и вычислительных ресурсов и, вследствие этого, их использование для постоянного (оперативного) наблюдения за высокой динамикой современного интернета представляется довольно проблематичным.

2 Идея исследования

В основу подхода, предложенного авторами настоящей статьи для определения размеров и качества индексов поисковых систем и объема русскоязычной части веба (Рунета), положены следующие ключевые идеи:

а) общая методика сравнения размеров индексов поисковиков и объема интернета примерно соответствует подходам, изложенным в обзорной части статьи (см. работы [2, 3, 4]) и модифицированным с учетом сегодняшних реалий.

б) для статистически обоснованного сэмплирования (получения достаточно надежной репрезентативной выборки) архивов поисковых машин можно использовать анализ выдач поисковиков, сделанных в ответ на сравнительно небольшой (порядка 100-200 элементов) массив редких однословных запросов (для которых объем выдачи в результатах поиска системы Яндекс не превышает, например, 500-600 документов). Предполагается, что в этом случае поисковая система отключает какие-либо дополнительные фильтры и выдает действительно полный набор документов из своего архива, содержащих искомые слова.

в) использование вышеуказанного массива для формирования русскоязычных запросов к преимущественно англоязычным поисковикам (типа Google или Yahoo!) вполне допустимо, и анализ полученных ответов статистически корректно отражает именно русскоязычную часть корпуса веб-страниц этих поисковиков.

г) имеется приблизительное равенство отношений объемов результатов сэмплирования каких-либо двух поисковых машин и отношений реальных объемов их архивов, т.е. существует примерно пропорциональная зависимость между изменением (увеличением) объема всего архива после очередного цикла обхода робота и той его частью, которая содержит документы с редкими словами. Данная гипотеза опирается на многолетний опыт разработки и эксплуатации поисковой системы Яндекс.

д) для комплексной оценки архивов поисковых систем необходимы не только «количественные» показатели (их размеры), но также и «качественные», такие как «чистота» и «свежесть»

документальных баз, т.е. доли полудубликатов и устаревших ссылок в них.

е) мониторинг за состоянием архивов поисковых машин должен быть полностью автоматическим, выполняться на регулярной основе и занимать разумное время, не превышающее нескольких часов.

3 Описание методов, алгоритмов и экспериментов

3.1 Определение относительных размеров архивов поисковых систем

Как уже указывалось выше, основным методом оценки размеров и качества баз данных является анализ результатов серий запросов к поисковикам, выполненных через общедоступный интерфейс. Каждая серия состоит примерно из 120 однословных запросов, использующих несколько «экзотические» редкие слова русского языка. Фрагмент такого списка слов, упорядоченный по убыванию «редкости» приведен ниже в Таблице 1, а полный список в алфавитном порядке - в Приложении 1.

Таблица 1

Фрагмент списка редких слов русского языка, используемых для запросов к поисковым машинам

поднормаль	мастодонзавр
пилохвост	мегаэволюция
плеврококк	паркеризация
орисфера	зубанка
логарифмика	координатомер
волкозуб	интразональность
базификация	диптанк
многозуб	модфа
камералистика	панзоотия
кальцифир	инкреция

Выбор слов определялся единственным условием, чтобы число документов в архиве Яндекса, содержащих искомое слово, не превышало некоторого небольшого порога (в экспериментах использовалось значение 500). Данное требование опирается на вполне разумную гипотезу, что при наличии в базе данных небольшого количества документов, релевантных запросу, поисковые системы не будут использовать какие-либо ограничительные фильтры, а выдадут все имеющиеся документы пользователю, хотя, возможно, в различном порядке, учитывая специфику алгоритмов ранжирования.

Для статистического обоснования размера серии (120 запросов) были проведены многочисленные эксперименты с сериями различного объема: от 30 до 200 запросов. Главный итог этих экспериментов – практически неизменные результаты после достижения размера 60-70 запросов. Поэтому для сохранния разумного баланса между временем проведения расчетов и исключением различных случайностей, могущих сильно исказить результаты, и был выбран удвоенный «запас прочности».

Для удобства автоматической выдачи запросов поисковикам и анализа их ответов, для каждой машины был создан специальный конфигурационный файл, содержащий всю необходимую для обработки информацию и состоящий из 5 основных элементов:

- URL начальной страницы поисковой системы
- шаблон общедоступного интерфейса для выполнения запроса
- регулярное выражение для распознавания URL'ов проиндексированных документов в результатах поиска
- регулярное выражение для распознавания ссылки «Еще с сайта»
- регулярное выражение для распознавания ссылки «Следующая страница»

В Таблице 2 приведены примеры конфигурационных файлов для поисковых систем Яндекс и Google, а их полный перечень находится в Приложении 2.

Таблица 2

Конфигурационные файлы для поисковых систем Яндекс и Google

Яндекс
http://www.yandex.ru
/yandsearch?pag=d&ag0=h&text=
web\/pos=\d+'\)" href="([^"]+)
"([^"]+)">Еще с сайта
<a href="(/yandpage[^"]+)" target=_self>следующая

Google
http://www.google.com
/search?hl=en&lr=lang_en|lang_ru&ie=CP1251&q=
<p class=g><a href=([^>]+)>
\[<a class=fl href=([^>]+)>More results from
<a href=([^>]+)><img src=/nav_next

Если поисковая машина возвращает результаты поиска в формате, отличающемся от наиболее популярной в Рунете кодировки windows-1251, то в конфигурационном файле дополнительно (через точку после имени поисковика) указывается тип кодировки (K - koi8-r, U - utf-8 и т.д.), например: «Рамблер.K», «Yahoo!.U» (см. также Приложение 2). В этом случае результаты выдачи перед обработкой перекодируются в windows-1251.

После получения ответа на запрос, с помощью конфигурационного файла выполняется обработка всех страниц, возвращенных поисковой системой, включая переходы по ссылке «Еще с сайта», и составляется полный список найденных URL'ов. В Приложении 3 в обобщенном виде (прототип

скрипта на Perl) приведена процедура такой обработки поисковой выдачи. Данный процесс повторяется для всех поисковиков и в результате мы получаем для каждого запроса общий список возвращенных всеми машинами URL'ов с указанием для каждого URL'а тех систем, в выдаче которых он присутствовал.

Затем, на основе полученного списка URL'ов, выполняется подкачка соответствующих документов из веба. Для каждого URL'а выполняется проверка на «свежесть» (см. ниже п.3.4) путем поиска вхождения основы словоформы запроса в текст соответствующего документа. Далее, загруженные документы кластеризуются на основе их сходства по содержанию в группы, содержащие нечеткие дубликаты. Результаты кластеризации используются в дальнейшем для определения «чистоты» архивов поисковых систем (см. ниже п.3.3).

После этого, для каждого поисковика подсчитывается суммарное количество найденных URL'ов по всем запросам (в качестве альтернативного варианта вместо URL'ов можно использовать построенные кластеры, поскольку, как показали проведенные дополнительные эксперименты, окончательные результаты практически совпадают – расхождения составляют не более 3-5%) и определяется отношение полученного числа к числу URL'ов, найденных Яндексом. Предполагается, что полученные дроби дают нижнюю оценку размеров русскоязычной части индексов поисковых машин по отношению к размерам базы Яндекса (см. [2]).

Для определения абсолютных размеров архивных хранилищ достаточно умножить относительные размеры поисковых машин, о которых говорилось в предыдущем абзаце, на абсолютный размер базы Яндекса, сведения о котором систематически публикуются в открытом доступе по адресу http://company.yandex.ru/ и который, например, на момент написания статьи (апрель 2006 года) был равен 852,643,000 документам (см. ниже п.3.5).

3.2 Определение объема Рунета (русскоязычной части веба)

Общий список возвращенных поисковиками URL'ов, может быть также использован для определения доли Яндекса (или других машин) в выдаче остальных поисковиков. Если эти доли примерно одинаковы, то в соответствии с представлениями теории вероятностей о независимых событиях и условных вероятностях можно считать, что алгоритмы построения индексов поисковыми системами также являются независимыми и, следовательно, доля Яндекса по отношению к объему Рунета приблизительно равна его доле в выдачах других поисковиков (см. также [2]). Тогда объем Рунета можно легко определить путем деления числа документов в базе Яндекса (см. пред. раздел) на величину этой доли. Для дополнительного повышения точности оценки можно использовать среднее арифметическое этих долей по всем исследуемым поисковикам.

Как показали проведенные эксперименты (см. ниже п.3.5), гипотеза о независимости алгоритмов индексирования различных поисковых машин достаточно надежно подтверждается и, следовательно, предложенный метод оценки вполне приемлем с точки зрения достоверности полученных результатов.

3.3 Определение показателя «чистоты» базы поисковых машин

Показатель «чистоты» отражает долю полудубликатов в архивах поисковых машин и определяется на основе кластеризации документов, загруженных из веба по результатам поиска, полученным в ответ на серию специальным образом подобранных однословных запросов, состоящих из редких слов (см. выше).

Для кластеризации используется наиболее простой и быстрый агломеративный метод одиночной связи, при котором документ присоединяется к уже существующему кластеру, если по крайней мере один из документов кластера имеет уровень сходства с включаемым документом, превышающий некоторый порог.

Для определения сходства между веб-документами используются два вида сигнатур: хеш-код документа (128-битная «дактилограмма» MD5) и логарифмический набор шинглов (оригинальная авторская модификация классического подхода, изложенного в [5]), т.е. контрольных сумм (типа CRC32), вычисляемых для непрерывных пятисловных последовательностей. При этом полученное числовое значение суммы должно делиться на некоторую минимальную степень двойки. После этого в результирующий набор включаются наименьшие значения из групп, образованных шинглами, дающими одинаковые остатки при делении на последовательные степени двойки вплоть до некоторой максимальной степени.

Документы считаются «похожими» по содержанию, если у них совпадают хеш-коды или значение «Dice coefficient» наборов шинглов не ниже некоторого порога (в нашем случае 0.75). Для повышения точности кластеризации у «похожих» (в смысле «Dice coefficient») документов дополнительно проверяется метрика Левенштейна (расстояние редактирования) с минимальным пороговым значением 0.85.

Показатель «чистоты» базы определяется как отношение числа кластеров, построенных (на основе введенной выше метрики сходства) для массива реальных документов, загруженных из веба в ответ на все однословные запросы, сделанные к определенной поисковой системе, к общему числу этих документов в массиве.

Данный показатель характеризует, в первую очередь, качество архивных хранилищ, так как наличие в поисковой выдаче высокого процента

«похожих» документов приводит к снижению релевантности результатов поиска.

3.4 Определение показателя «свежести» индекса поисковых систем

Показатель «свежести» отражает долю устаревших ссылок в индексе поисковых машин, т.е. процент документов, изменивших свое содержание с момента последнего обхода робота или вообще больше не существующих (по крайней мере по URL-адресу, хранимому в поисковой системе). Для определения таких документов делается попытка найти псевдоосновы (а не сами словоформы - для увеличения точности результатов) исходных слов запросов, получающиеся путем отбрасывания двух последних букв в текстах документов, загруженных из веба. При неудачном поиске документ считается измененным, а ссылка на него в архиве - устаревшей.

Коэффициент «свежести» архива равен дроби, в знаменателе которой находится общее число элементов в списке URL'ов, возвращенных какой-либо поисковой машиной в ответ на все сделанные редкие запросы, а в числителе - число документов, скачанных из интернета на основе этого списка URL'ов и содержащих псевдоосновы словоформ этих запросов.

Данный показатель является важным параметром, определяющим (как и показатель «чистоты» - см. пред. раздел) качество поисковой системы, поскольку, если в результатах поиска присутствует большое количество устаревших ссылок, выдачу вряд ли можно назвать релевантной.

3.5 Описание экспериментов и анализ полученных результатов

Систематические эксперименты в Яндексе по измерению размеров архивов поисковых машин и определению объема Рунета начались в феврале 2004 года и продолжаются вплоть до сегодняшнего дня (апрель 2006 года) с интервалом 1-2 недели. В этих экспериментах, в различное время, участвовали все наиболее крупные российские и зарубежные поисковики, такие как Яндекс, Рамблер, Апорт, Google, Yahoo!, AltaVista, Fast, Msn и др.

Как уже указывалось выше, в процессе эксперимента к каждой поисковой машине посылается серия примерно из 120 редких однословных запросов и полученные результаты поиска анализируются с целью определения попарного пересечения выдач и вычисления показателей «чистоты» и «свежести».

Общий объем всего списка URL'ов, получаемых от всех поисковиков, изменялся за время наблюдений в диапазоне от 20 тыс. до 40 тыс. элементов. Полное число URL'ов, выдаваемых в ответ на один запрос составляло в среднем от 100 до 300 (полный разброс - от 50 до 500).

Поскольку зарубежные поисковые системы, как правило, не имеют средств морфологического анализа русских словоформ, то для обеспечения статистической корректности условий эксперимента, запросы к Яндексу выполнялись дважды - в обычном режиме и режиме с отключенной морфологией.

В результате экспериментов было установлено, что доля Яндекса в выдаче других поисковых машин оставалась практически постоянной во время каждого отдельного эксперимента и менялась весьма незначительно от одного эксперимента к другому, в пределах от 0.32 до 0.36 за все время наблюдений. Это подтверждает гипотезу о независимости алгоритма индексирования веба Яндексом от других поисковиков и дает уверенность в надежности полученных оценок объема Рунета, приведенных в Таблице 3.

Таблица 3

Динамика изменения объема Рунета в веб-документах

Февраль 2004	428,571,000
Март 2004	504,658,000
Октябрь 2004	747,503,000
Ноябрь 2004	799,003,000
Апрель 2005	1,555,937,000
Май 2005	2,020,463,000
Июль 2005	2,101,304,000
Август 2005	2,191,782,000
Сентябрь 2005	2,355,390,000
Декабрь 2005	2,538,893,000

Данные расчеты выполнялись для множества различных серий запросов и, поскольку окончательные результаты практически полностью совпадали, вполне естественно было ограничиться точечными оценками и не использовать метод доверительных интервалов.

Полученные данные о величине объема Рунета нашли дополнительное независимое подтверждение при анализе базы URL'ов, известных роботу Яндекса. На сегодняшний день размер этой базы примерно равен 3.3 млрд., из которых около 900 млн. являются дубликатами, поэтому общий размер базы составляет порядка 2.4 млрд. URL'ов. В этот список не включены «специальные» URL'ы, адресующие скрипты, и «простые» URL'ы, содержащие менее 3-х слэшей. Общее число таких URL'ов - приблизительно 10% от всей базы, т.е. около 0.24 млрд. Таким образом, окончательная оценка получается примерно 2.64 млрд. документов, что достаточно близко к вышеприведенным результатам, полученным статистическими методами сэмплирования архивов поисковых систем.

В таблице 4 представлены сведения о размерах баз данных и показателях «чистоты» и «свежести» наиболее крупных российских и зарубежных поисковых машин, полученные в конце 2005 года.

Таблица 4

Размеры индексов и показатели качества архивов крупнейших поисковых систем Рунета

Система	Размер	Чистота	Свежесть
Яндекс	828,991,352	0.901	0.839
Рамблер	1,019,659,362	0.759	0.801
Google	547,134,292	0.884	0.806
Yahoo!	406,205,762	0.907	0.904
MSN	348,176,367	0.950	0.920

3.6 Программная реализация метода

Изложенные выше подходы и алгоритмы реализованы в виде четырех программных модулей, написанных на языке Perl. Общий объем всех этих модулей - порядка 800 операторов языка.

Запуск модулей выполняется автоматически и на регулярной основе с интервалом 1 раз в неделю. Время расчетов составляет примерно 10-12 часов на обычной десктопной машине средней производительности.

4. Выводы и обсуждение результатов

В настоящей работе была сделана первая попытка оценить количественные и качественные характеристики русскоязычных частей индексов современных российских и зарубежных поисковых машин и определить размер и темпы роста Рунета - стремительно развивающегося сегмента Сети.

В основе подхода лежали идеи, впервые высказанные в исследованиях [2, 3, 4] и развитые и дополненные авторами настоящей статьи применительно к задачам, перечисленным в разд.2.

В работе обобщены данные экспериментов, проводимых в течение последних двух лет в компании Яндекс по наблюдению за состоянием и динамикой архивных хранилищ наиболее крупных поисковых систем и русскоязычной частью веба.

В результате этих экспериментов было установлено, что объем Рунета на конец 2005 года составляет порядка 2.5 млрд. документов. Размеры баз данных крупнейших российских поисковиков Яндекса и Рамблера, с учетом качественных показателей «чистоты» и «свежести», примерно равны (около 1 млрд. документов) и в 2-2.5 раза превосходят русскоязычные части корпусов веб-страниц наиболее популярных западных поисковых машин Google, Yahoo! и Msn, объемы которых тоже примерно равны между собой.

Использованные методы сэмплирования документальных архивов основаны на применении редких однословных запросов к поисковикам, что позволило избежать основных недостатков, присущих большинству проанализированных выше зарубежных аналогов - смещенности выборок в сторону длинных и высокоранжированных документов. Для борьбы с этими недостатками, как правило, требуются значительные усложнения используемых алгоритмов и большие временные и ресурсные затраты [6]. В нашем же случае выбираются все имеющиеся в базах поисковиков документы, хотя, возможно, в различном порядке, учитывая специфику алгоритмов ранжирования.

Литература

[1] D. Brake. Lost in Cyberspace. New Scientist, June 28, 1997.

[2] K. Bharat and A. Broder. A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. Proc. of the 7th International World Wide Web Conference, April 1998.

[3] K. Bharat and A. Broder. Measuring the Web. March 1998. http://www.research.digital.com/SRC/whatsnew/sem.html

[4] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science, vol. 280, no. 5360, April 3, 1998.

[5] D. Brake, A. Broder, S. Glassman, M. Manasse and G. Zweig. Syntactic clustering of the Web. Proc. of the 6th International World Wide Web Conference, April 1997.

[6] Z. Bar-Yossef, M. Gurevich. Random Sampling from a Search Engine's Index. . Proc. of the 15th International World Wide Web Conference, May 2006.

Приложение 1. Список редких слов русского языка, используемых для запросов к поисковым машинам

абхидхарма
автопрокладчик
авункулат
агитфильм
агробиоценоз
аквафортист
амфотерность
анахоретство
антипассат
апофема
аргонавтика
базед
базификация
базофилия
баллистит
баллистокардиография
безлопастный
бекасинник
белотал
бладшот
блокшив
брахиозавр
брекватер
бриолин
бхакти
вантоз
веберметр
верцеф

виверра
вишнуизм
внеиндивид-ый
водоемкость
водорез
волкозуб
вороновица
газообильность
галломания
гидромедуза
гименофор
гипсометрия
гнюс
декатировка
диптанк
зоохория
зубанка
иглофильтр
иллюминирование
инклинометр
инкреция
интразональность
иподьякон
калуфер
кальцифир
камедетечение
камералистика
канефора
квадруполь
квартоль
коконосушилка
компрачикосы
контроверсия
координатомер
кроссбридинг
литорина
логарифмика
лополит
луншань
мастодонзавр
мегаэволюция
металлооптика
многозуб
многоцветница
модфа
монохромия
нейросекреция
неликвидность
несебыр
нижнеянск
нонет
нунатак
обвойник
одревеснение
ольвиополь
оптиметр
орисфера
оркестрион
ортит
ослинник
палинодия
панзоотия
панмиксия
паралингвистика
парапроцесс
паратаксис
паринирвана
паркеризация
паропреобразователь
перверсия
перловник
пилохвост
пиранометр
плеврококк
поднормаль
подорешник
подчашие
пчелоопыление
резольвента
рельсоукладчик
сансевьера
сантистокс
селенография
сигариллы
слепоглухонемота
спелеометрия
фаготерапия
фазорегулятор
хлораль
цивета
цинкит
цхалтубо
эвглена
ямокопатель

Приложение 2. Конфигурационные файлы для наиболее крупных систем Рунета

Яндекс
http://www.yandex.ru
/yandsearch?pag=d&ag0=h&text=
web\/pos=\d+'\)" href="([^"]+)
"([^"]+)">Еще с сайта
<a href="(/yandpage[^"]+)" target=_self>следующая

Google
http://www.google.com
/search?hl=en&lr=lang_en|lang_ru&ie=CP1251&q=
<p class=g><a href=([^>]+)>
\[<a class=fl href=([^>]+)>More results from
<a href=([^>]+)><img src=/nav_next

Рамблер.К
http://search.rambler.ru
/srch?words=
class="ttl">.+?href="([^"]+)"
\(<a href="([^"]+)">всего
<a href="([^"]+)">следующая

Yahoo!.U
http://search.yahoo.com
/search?ei=UTF-8&fr=sfp&p=
<a class=yschttl.+**(http%3A%2F%2F[^"]+)
<a href="(.+)">More from this site
<a href="(.+)">Next

Приложение 3. Обобщенная процедура обработки поисковой выдачи

```
# Выполняется для всех поисковиков в цикле для одного запроса
# сформировать поисковый url
$url = $search.uri_escape($query);
# повторять
do {
   # выполнить запрос и скачать ответ
   return if (($response = download($url)) eq "");
   # извлечь и добавить все "найденные документы" из страницы ответа
   get_docs($response, $docs);
   # если ссылки есть "еще с сервера" то для всех таких ссылок повторять
   while ($response =~ m!$reSiteMore!g) {
   # сформировать url запроса "еще с сервера"
   $more = $1;
      # повторять
   do {
      # выполнить запрос и скачать ответ
      next if (($more_response = download($more)) eq "");
      # извлечь и добавить все "документы" из страницы ответа
      get_docs($more_response, $docs);
      # пока можно найти и сформировать url перехода на следующую страницу
   } while (($more) = ($more_response =~ m!$reNextPage!));
   }
# пока можно найти и сформировать url перехода на следующую страницу
} while (($url) = ($response =~ m!$reNextPage!));
```

Methods for Comparative Analysis of Modern Search Systems and Runet Size Determination

© Segalovich Ilya V., Zelenkov Yuri. G., Nagornov Denis O.
Yandex, Moscow

In the presented work automatic methods for comparing the qualitative and the quantitative characteristics of the Russian-language part of indices of contemporary search systems and for determining the Runet size are considered.

The actuality of the research is determined by extremely high rates of increase in the Russian-language part of the Web. Thus, according to the results obtained by the authors in February 2004 (the beginning of systematic observations), the Runet size was about 400 million documents, and at the end of 2005 it was about 2500 million. Therefore, it increased more than six-fold (!) for 2 years. In accordance with the latest measurements, this rate will possibly remain the same in the near future.

The novelty of the work is determined by the following:

1. There are no similar works connected with the Russian-language part of the Internet;
2. In the given work not only the quantitative parameters but also the qualitative parameters (“cleanliness” and “freshness”) of archives of the largest search systems of the Internet such as Yandex, Rambler, Google, and Yahoo! are considered.

Furthermore, by comparison with the foreign analogues, the methods developed by the authors allow you to make daily measurements of the Web. This is especially important taking into account very high dynamics of changes in the modern Internet.