Структура русскоязычной части глубинного Веба*

Денис Шестаков[1], Наталья Воронцова[2]

[1] University of Turku, Finland, da_shestakov@yahoo.com
[2] Московская Медицинская Академия им. И.М.Сеченова, n_vorontsova@myrealbox.com

Аннотация

Значительная часть Веба, «скрытая» за поисковыми веб-формами, не индексируется современными поисковыми системами. «Скрытая» часть Веба известна под названием глубинный Веб. Веб-страницы в глубинном Вебе являются динамическими и представляют собой результаты, возвращаемые поисковыми веб-формами. В нашей работе, мы изучали русскоязычную часть глубинного Веба (RDW). Основной целью было определение числа глубинных сайтов, расположенных в RDW. Данное исследование является первой работой, которая рассматривает определенную часть глубинного Веба, представляющую интерес носителям какого-то одного (не английского) языка.

1. Введение

Благодаря развитию веб-технологий в настоящее время Веб представлен страницами, как со статичным, так и с динамичным содержанием. Динамичность ведет к большей интерактивности веб-страниц, но, в тоже время, к тому, что многие динамические страницы не индексируются современными поисковыми системами (например, Google.com или Yandex.ru). Поисковые системы индексируют и, значит, позволяют производить поиск только среди «индексируемой» части Веба, известной как *индексируемый Веб (publicly indexable Web)* [13]. В индексируемый Веб, в частности, не входят ог-

* Авторы выражают признательность компании «Яндекс» за проведение конкурса научных стипендий, за предоставленную стипендию (заявка 102104) и за доступ к данным.

ромное количество веб-страниц, возвращаемых поисковыми веб-формами.

Значительная часть Веба «скрыта» за поисковыми веб-формами, которые являются единственными веб-интерфейсами к сотням тысяч (около 450 000 согласно [5]) баз данных доступных онлайн. «Скрытая» (находящаяся за поисковыми веб-интерфейсами) часть Веба известна как минимум под тремя названиями: *скрытый Веб (hidden Web)* [8][16], *глубинный Веб (**d**eep **W**eb)* [2][18] (используемое в данной работе) и *невидимый Веб (invisible Web)* [17]. Веб-страницы в глубинном Вебе (далее сокращенно **DW**) являются динамическими[1] и представляют собой результаты, возвращаемые поисковыми веб-формами. DW является очевидным и интересным объектом для исследования, так как для большинства веб-пользователей поиск нужной информации в Вебе основывается исключительно на результатах, возвращаемых той или иной поисковой системой. В результате игнорируется огромное число веб-документов из DW.

В данной работе мы изучали русскоязычную часть глубинного Веба (далее сокращенно **RDW**). Нашей целью было количественное определение основных параметров RDW. Мы полагаем, что полученные характеристики вместе с кратким введением в DW будут интересны широкому кругу русскоязычных читателей. Помимо этого, мы составили небольшой список русскоязычных ресурсов DW, который может использоваться (и дополняться) при дальнейшем изучении RDW/DW.

Насколько мы знаем, данное исследование является первой работой, которая рассматривает лишь определенную часть DW – а именно часть, которая представляет интерес преимущественно носителям какого-то одного (не английского) языка. Выбор именно русскоязычной части DW объясняется тем, что для обоих авторов русский является родным языком, а также наличием доступа к данным, которые относятся к русскоязычному Вебу и были предоставлены русскоязычной поисковой системой Яндекс.

Наш отчет состоит из несколько частей. В следующем разделе мы дадим более подробное описание понятия «глубинный Веб». Далее, мы представим обзор литературы, посвященной измерению характеристик DW и извлечению информации из DW. Раздел 4 расскажет об определении основных характеристик DW. Описание экспериментов и результатов дано в разделе 5. Разделы 6 и 7 представят нашу интерпретацию полученных результатов и заключение соответственно.

2. Глубинный Веб

Глубинный Веб (DW) определяется как множество динамически генерируемых веб-страниц, содержащих данные из баз данных доступных онлайн. Веб-страницы, наполненные информацией из баз данных (*ресурсов DW*), также известны как *бд-наполненные страницы (data-rich/data-intensive pages)*. Удобно выделять три следующие понятия: *глубинный сайт* – веб-сайт/сервер, предоставляющий информацию из одной или нескольких *онлайн баз данных (searchable/web database)*, каждая из которых доступна через один или несколько *веб-интерфейсов (query interface)*. Таким образом, наличие, как минимум, одного веб-интерфейса к базе данных превращает обычный веб-сайт в глубинный. В настоящее время, в большинстве случаев веб-интерфейсы к онлайн базам данных являются поисковыми HTML формами, поэтому часто отождествляют веб-интерфейс с поисковой HTML формой[2].

Мы склонны ставить знак равенства между глубинным сайтом и онлайн базой данных, так как это упрощает изучение ресурсов DW. Согласно нашей точки зрения, глубинный сайт ведет к двум и более базам данных, если веб-интерфейсы, расположенные на сайте, позволяют производить поиск среди объектов, задаваемых двумя и более сильно отличающимися друг от друга структурами. Например, глубинный сайт с двумя веб-формами, предназначенными для поиска автомобилей и авто-деталей, будет рассматриваться нами, как сайт с двумя базами данных. В тоже время, сайт с поиском по автомобилям и мотоциклам (и то и то автотранспортные средства), как сайт с одной базой данных.

Веб-формы предназначены для использования людьми, а не компьютерными приложениями или программными агентами. В силу этого у поисковых систем возникают трудности с индексацией веб-страниц, возвращаемых как результаты запросов через веб-формы к базам данных. Другими словами, информация из онлайн баз данных «скрыта» за поисковыми веб-формами и, таким образом, «скрыта» или «невидима» для поисковых систем[3]. Здесь следует подчеркнуть, что сами страницы с поисковыми веб-формами проиндексированы и известны поисковикам[4]. Так, под русскоязычной частью DW мы понимаем множество бд-наполненных страниц, к которым можно придти через веб-интерфейсы, известные русскоязычной поисковой системе Yandex.ru. Для объективности отметим, что DW частично индексируется поисковиками, причем для отдельных глубинных сайтов может индексироваться более, чем 80% всех динамически генерируемых страниц.

Веб-форма является основным индикатором наличия (или отсутствия) за ней онлайн базы данных, а также указывает на тематику базы данных. Формы для навигации по сайту, авторизации, регистрации, подписки, голосования, отправления сообщений, а также формы, предоставляющие поиск по сайту, исключаются, так как первые не веб-интерфейсы к базам данных, а последние производят поиск по веб-страницам из индексируемого Веба. Помимо этого, мы исключили из DW содержимое поисковых систем. Также, мы не рассматривали новостные сайты (которые часто предоставляют поиск по архиву своих новостей), увидев, что в настоящее время содержимое новостных сайтов хорошо покрывается поисковыми системами. В тоже самое время, большинство веб-форумов рассматривалось нами как часть DW (есть база данных сообщений и возможность поиска по ним). В случаях, когда не было ни одного веб-интерфейса, но было очевидно наличие базы данных, наполняющей сайт, мы считали, что сайт глубинный, если поисковики индексировали менее трети веб-страниц, генерируемых на сайте.

Рисунок 1: Веб-форма для поиска автомобилей

С точки зрения большинства веб-пользователей поисковые системы

«знают» если и не обо всей информации в Вебе, то, по крайней мере, о ее большей и наиболее существенной части. Это значит, что при поиске в Вебе, запрос к поисковику часто строится таким образом, чтобы возвращаемые ссылки сразу вели к страницам с нужной информацией. С другой стороны, пользователь, осведомленный о существовании DW, может улучшить/дополнить свой поиск, если сначала найдет с помощью поисковой системы (или сервиса по поиску ресурсов DW) один или более веб-интерфейсов к интересующим его базам данных, а затем, составив с помощью найденных интерфейсов более специализированные запросы, найдет требуемую информацию.

Для иллюстрации вышесказанного, рассмотрим глубинный сайт Auto.ru, содержащий информацию об автомобилях в России. На сайте размещены множество веб-форм, позволяющих искать предложения о продаже новых и подержанных автомобилей. На рисунке 1 показана одна из таких веб-форм (ее адрес – http://www.cars.auto.ru/find) – она предназначена для поиска подержанных автомобилей любой марки.

Результаты заполнения и отправки формы – одна или несколько веб-страниц, содержащая список ссылок-результатов. Каждая ссылка из списка ведет к веб-странице (бд-наполненной странице) с предложением, описывающим один автомобиль. Страницы с предложениями (имеющие адреса вида http://www.cars.auto.ru/sale/*) сгенерированы сервером на основе общего для всех таких страниц шаблона, где данные (о конкретном автомобиле), вставляемые сервером в шаблон, берутся из базы данных на Auto.ru. Воспользовавшись расширенным поиском Яндекса, нетрудно узнать, какая часть предложений о продаже поддержанных автомобилей (а их более **74000** на момент июня 2005г.) проиндексирована Yandex.ru. Яндекс (в июне 2005г.) знал только о **пяти (!)** таких страницах, причем один из результатов вел к пустой странице, т.е. к устаревшему предложению. Таким образом, мы видим, что информация из базы данных на Auto.ru (по крайней мере, из той ее части, что касается подержанных автомобилей) практически полностью скрыта от Яндекса.

3. Краткий обзор литературы о DW

Как уже было указано во введении, «неиндексируемая» «скрытая» часть Веба в литературе известна под следующими названиями: скрытый Веб (hidden Web – [8][16]), глубинный Веб (deep Web - [2][18]), и невидимый Веб (invisible Web - [17]). В данном обзоре мы, во-первых, укажем на две ключевые работы об измерении ос-

новных характеристик DW, а, затем, кратко расскажем о круге проблем, возникающих при извлечении данных из DW.

Работа [2], являющаяся, вероятно, одной из наиболее известных работ о DW, содержит как введение в DW, так и численные характеристики DW. Несмотря на то, что она была подготовлена в качестве сопроводительного документа к программному продукту коммерческой компании (т.е. не может считаться полноценным научным исследованием), полученные в ней оценки DW довольно известны и цитируемы. Итак, приведем наиболее важные параметры DW на момент марта 2000г., полученные в [2]:

- DW представлен по меньшей мере 43 000-96 000 глубинных сайтов (причем итоговая оценка - более чем 200 000 глубинных сайтов).
- Размер DW превышает размер индексируемого Веба на два порядка (в 400-550 раз).
- 95% DW – это данные, к которым предоставлен свободный доступ (не требующий подписки или платы).

В настоящий момент мы ставим под сомнение оценку размера DW в 400-550 размеров индексируемого Веба. Согласно нашему анализу (неопубликованные данные), DW имеет существенно меньший размер, чем утверждается в [2]. С другой стороны, оценка количества глубинных сайтов в число порядка 10^5 представляется нам более адекватной. Также заметим, что метод, который использовался для определения числа баз данных (“overlap analysis”, см.[2]), не является идеальным, так как требует определенных допущений, которые, весьма вероятно, неверны для реальных данных.

Более свежие оценки характеристик DW были получены в работе [5]. В ней авторы поставили задачу оценить количество глубинных сайтов, онлайн баз данных и веб-интерфейсов во всем DW. Согласно их оценкам, которые на сегодняшний день наиболее достоверны и, что немаловажно, проверяемы, DW на момент апреля 2004г. представлен 300 000 глубинных сайтов, 450 000 онлайн баз данных и 1 250 000 веб-интерфейсов. В следующем разделе мы подробно остановимся на недостатках метода случайной выборки из списка IP-адресов (*random IP-sampling method*), который использовался для оценки числа глубинных сайтов. Сейчас же укажем на субъективность при определении количества онлайн баз данных и веб-интерфейсов. Дело в том, что определение как числа онлайн баз данных, находящихся на определенном глубинном сайте, так и числа веб-интерфейсов к базам данных, часто неоднозначно. Для примера рассмотрим глубинный сайт Auto.ru, который мы уже описывали в предыдущем разделе. Auto.ru позволяет искать предложения о

продаже новых (по адресу http://www.new.auto.ru/find) и подержанных автомобилей (http://www.cars.auto.ru/find). С одной стороны, можно утверждать, что поиск предложения о, скажем, новом автомобиле ведется именно в базе данных о новых автомобилях и, следовательно, на Auto.ru расположены, как минимум, две онлайн базы данных[5]. С другой стороны, вполне разумно предположить, что физически вся информация как о подержанных, так и о новых автомобилях содержится только в одной базе данных[6].

Другой интересный результат, представленный в [5], - оценка доли DW, индексируемой поисковыми системами. Рассмотрев 80 ресурсов DW различной тематики, Chang et al. получили, что поисковая система (рассматривался Google.com) покрывает индексом около 25% бд-наполненных страниц[7]. Действительно, если глубинный сайт помимо поисковых форм также имеет средство навигации по содержимому онлайн базы данных (обычно в виде каталога или классификатора), то, значит, к бд-наполненным страницам ведут прямые ссылки. Это, в свою очередь, позволяет поисковикам индексировать эти динамические веб-страницы. Наличие навигации на глубинном сайте вовсе не редкое явление и, более того, очень типично для сайтов с такой тематикой как книги, музыка или фильмы (у не менее 70% рассмотренных в [5] глубинных сайтов с такими тематиками есть навигация по содержимому баз данных). Работа [6] описывает автоматическую систему, позволяющую извлекать данные с сайтов, содержимое которых представлено в виде каталога.

Различным аспектам извлечения и интеграции данных из DW посвящено довольно много исследований. В данном обзоре мы расскажем лишь о нескольких из них. Более детальный обзор подходов/методов, используемых для извлечения/интеграции веб-данных (web data extraction/integration), может быть найден, например, в [8][12]. Дополнительно мы хотим отметить, что значительная часть проблем, возникающих при автоматическом извлечении/интеграции веб-данных, обусловлена языком разметки HTML. Причина здесь в том, что HTML-формат удобен для представления данных человеку, но не компьютерному приложению. Тем не менее, повсеместность HTML в Вебе не оставляет какого-либо выбора автоматическим (машинным) средствам извлечения/интеграции веб-данных.

Итак, для извлечения информации из онлайн базы данных нужно:

1) найти поисковый веб-интерфейс к интересующей базе данных;
2) с помощью интерфейса сформулировать и отправить запрос к базе данных;
3) получить и обработать ответ сервера с результатами запро-

са, а именно: определить по ответу, не было ли в запросе ошибок и удачен ли он (возвращает ли результаты); для удачных запросов выполнить навигацию по результатам, чтобы извлечь все бд-наполненные страницы;

4) проанализировать страницы с результатами и извлечь из них данные (соответствующие записям базы данных) [4].

Шаг 1 (поиск онлайн базы данных) может быть осуществлен с помощью существующих директорий ресурсов DW [3][10] или с помощью общеизвестных поисковых систем.

При выполнении шага 2 (построение и отправка запроса) возникает ряд трудностей. Ключевой момент – идентификация *описателей полей веб-формы (form field labels)* (слово или фраза, указывающие пользователю, что должно быть введено в соответствующее поле формы). Например, для веб-формы на рис.1, описатель «Цвет» соответствует верхнему правому полю, а описатель «Год вып.» соответствует двум верхним полям слева. Несмотря на очевидные для человека, просматривающего веб-форму в окне браузера, соответствия между описателями и полями, установление таких же соответствий на основе HTML-кода (а именно его приходится анализировать программному агенту) является сложной задачей. Работа [11] рассказывает о подходах, которые можно использовать для определения описателей полей, а [16] предложит более точный метод. Описатели полей указывают на домены значений, которые могут быть введены в соответствующие поля. Домены значений должны быть в наличии и могут дополняться в процессе обработки результатов других запросов [16]. Другое затруднение вызвано случаями *последовательных форм (consecutive forms)*, т.е. когда для получения результатов требуется заполнить и отправить две и более веб-форм [18].

Работа [16] описывает поисковик (выполняющую шаги 2 и 3), способную индексировать веб-страницы, возвращаемые веб-формами. Один из основных недостатков системы в том, что она работает с очень простыми формами. Работы [1] [18] описывают подходы, которые могут использоваться для построения специализированных систем (выполняющих шаги 2, 3 и 4). Предлагаемые в этих работах системы обладают гибкими языками запросов, которые, в частности, позволяют извлекать данные из бд-наполненных страниц, использовать извлеченные данные для заполнения других веб-форм и сохранять данные в реляционной базе данных или в XML-формате.

К сожалению, мы не знакомы с оригинальными русскоязычными работами о DW или о методах работы с данными из DW. Выполнив поиск в Яндексе по словосочетаниям «скрытый Веб» и «невидимый Веб», мы увидели, что ряд авторов (как научных работ, так и статей

в популярных компьютерных журналах) знают о существовании DW и сопутствующих ему проблемах и задачах. Тем не менее, просмотренные нами статьи либо не делают своего вклада в изучение DW, либо имеют перед собой иные объекты исследования.

4. Определение основных характеристик DW

Одной из важнейших характеристик DW (и его части, RDW) является количество онлайн баз данных. Другой интересный, но менее объективный параметр - размер DW (суммарный размер бд-наполненных страниц).

4.1. Количество онлайн баз данных

Для измерения числа баз данных мы использовали метод *случайной выборки из списка хостов (random host-sampling)*. Мы решили использовать этот метод, так как:

1) нашли метод *случайной выборки из списка IP-адресов (random IP-sampling)*, описанный в работе [5], неоптимальным для данной задачи;
2) обладали репрезентативным списком хостов, предоставленным Яндексом (набор данных «Хостграф»).

Базовая идея обоих вышеупомянутых методов очень проста: из большого списка (IP-адресов или хостов), покрывающего изучаемый объект (например, весь Веб или Рунет, как в нашей работе), делается случайная выборка определенного объема, которая затем изучается. Полученные характеристики выборки можно перенести и на весь список (IP-адресов или хостов), а, значит, и на изучаемый объект, если: объем выборки достаточно большой; список полностью покрывает изучаемый объект; и изучаемые характеристики однородно распределены по списку.

Начнем с критического рассмотрения метода случайной выборки из списка IP-адресов, который в работе [5] использовался для определения количества глубинных сайтов, онлайн баз данных и веб-интерфейсов. Для этого рассматривался список всех действительных IP-адресов[8], делалась выборка в один миллион адресов, к которой затем посылались HTTP-запросы (проверялся порт 80), чтобы определить находятся ли по этим адресам веб-сервера. Далее, с головных страниц, расположенных на найденных веб-серверах, скачивались (с глубиной 3) и анализировались на наличие веб-интерфейсов все веб-страницы, к которым ведут ссылки с головных страниц.

Легко увидеть следующие недостатки данного метода. Во-первых, некоторым веб-сайтам (например, Google.com, Rambler.ru, Rbc.ru)

может соответствовать несколько IP-адресов. Для данного метода это означает, что вероятность попадания одного из таких сайтов в случайную выборку выше, чем у остальных сайтов. Во-вторых, метод не принимает во внимание виртуальный хостинг, т.е. очень распространенную ситуацию, когда на одном веб-сервере, располагается несколько веб-сайтов. В результате, для каждого IP-адреса $X_1.X_2.X_3.X_4$ из выборки рассматривается лишь «сайт по умолчанию» (возвращаемый по ссылке http://$X_1.X_2.X_3.X_4$/) и игнорируются все остальные сайты, физически расположенные на том же IP-адресе. Наконец, не учитывается неравномерное распределение веб-серверов по всему диапазону IP-адресов. В частности, известно, что в каждой из сетей вида X.0.0.0/8, где $2<X<223$, располагается разное количество веб-серверов.

Метод случайной выборки из списка хостов свободен от части недостатков, возникающих при работе с IP-адресами. Например, проблема виртуального хостинга отпадает сама собой. Между тем, использование метода случайной выборки из списка хостов сопряжено с определенными трудностями. Основное препятствие – получение репрезентативного списка хостов, покрывающего весь Веб или какую-то определенную часть Веба (Рунета в нашем случае). Нам не известно о существовании такого рода списков в свободном доступе[9].

Другое затруднение (схожее с проблемой «один сайт-несколько IP-адресов») заключается в том, что к одному и тому же сайту (в том числе глубинному) может вести несколько доменов. Для примера, научная электронная библиотека eLibrary.ru имеет два адреса: http://elibrary.ru и http://e-library.ru. Возникающая здесь проблема еще и в субъективности идентификации глубинного сайта. Для иллюстрации, вновь обратимся к глубинному сайту Auto.ru, который мы описывали ранее. При обзоре работы [5] в разделе 3 мы уже указывали, на примере Auto.ru, на неоднозначность определения количества онлайн баз данных на глубинном сайте. Схожая неоднозначность существует и при решении, какой именно сайт предоставляет доступ к онлайн базе данных. Например, любой из таких сайтов как auto.ru, cars.auto.ru (поиск по подержанным автомобилям), new.auto.ru или newauto.ru (поиск по новым автомобилям), moto.auto.ru (по мотоциклам) и несколько других в домене auto.ru можно считать глубинным. В самом деле, с головной страницы любого из них можно попасть к веб-интерфейсам, ведущим к одной и той же базе данных[10], следуя не более, чем трем ссылкам. Нетрудно увидеть, что для метода случайной выборки из списка хостов ситуация «несколько хостов–одна база данных» будет вести к завышению

количества онлайн баз данных.
Резюмируя, скажем, что метод выборки из хостов (при наличии репрезентативного списка хостов) должен давать более близкую к действительности оценку числа ресурсов DW, чем метод выборки из IP-адресов.

4.2. Размер DW

Оценка размера DW представляется весьма интересной задачей, так как позволяет сравнить объем информации в DW и в индексируемом Вебе. Однако существует несколько замечаний, понижающих роль данного параметра. Прежде всего, под размером DW будем понимать суммарный размер всех ресурсов DW, где размер ресурса DW – суммарный размер определенного набора бд-наполненных страниц, каждая из которых соответствует одной записи или одной логической сущности в онлайн базе данных. В свою очередь, размер бд-наполненной страницы вычисляется также, как и поисковыми системами: по размеру HTML-кода страницы, при этом размер графических изображений на странице не учитывается.

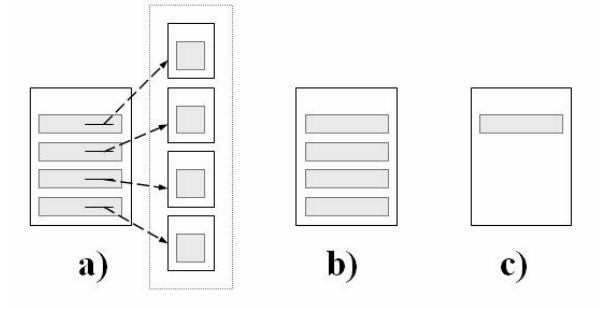

Рисунок 2: Три вида бд-наполненных страниц

Для большой наглядности обратимся к рисунку 2, где схематично изображены три вида бд-наполненных страниц, возвращаемых как результаты заполнения веб-форм:

a) Результат запроса - веб-страница (или несколько страниц, связанные навигационными ссылками), содержащая(-ие) «заголовок» записи (т.е. запись приведена не полностью, указаны ее основные параметры). При этом каждый заголовок включает ссылку, ведущую к странице или к файлу[11], где запись приводится в полном объеме. Например, веб-формы на Auto.ru возвращают результаты именно в таком

виде.

b) Возвращаемый результат – веб-страница (или несколько страниц, связанные навигационными ссылками), содержащая(-ие) более, чем одну запись. При этом каждая запись приводится полностью. Часто, такие страницы содержат информацию о каком-то объекте, где записи являются составными частями этого объекта. Например, для базы данных музыкальных файлов песни (записи в базе данных) группируются по музыкальным альбомам.

c) Результат запроса – веб-страница с одной записью. Типично для ресурсов DW, предоставляющих справочную информацию (в частности, для онлайн словарей и энциклопедий).

Как мы видим, в случаях a) и c) для оценки размера ресурса DW достаточно знать общее число записей в базе данных и средний размер бд-наполненной страницы (или файла), содержащей одну запись. Случай b) более сложный, так как для корректной оценки вначале необходимо определить логическую структуру данных. Например, для получения оценки размера ресурса DW, посвященного mp3-файлам, удобно определить общее число музыкальных альбомов, а затем найти средний размер бд-наполненной страницы, представляющей один альбом.

Нужно заметить, что определив размер того или иного ресурса DW, мы не получаем какую-либо новую информацию. Фактический размер базы данных будет в несколько раз меньше полученной оценки, потому что не учитывается избыточная HTML-разметка. Впрочем, и фактический размер базы данных мало, что скажет большинству веб-пользователей. В тоже время, число записей в базе данных представляется нам более важным и актуальным параметром, так как непосредственно указывает и на количество объектов, доступных для поиска, и на примерный размер базы данных.

5. Эксперименты

В июне 2005г. мы провели несколько экспериментов. Мы определяли число ресурсов RDW с помощью методов случайной выборки из списка IP-адресов и случайной выборки из списка хостов. В разделе 4.1 мы указывали на слабые стороны метода выборки из IP-адресов. Тем не менее, результаты, полученные с помощью этого метода, также представляют интерес, так как дают заниженную оценку числа ресурсов RDW.

5.1. Определение количества ресурсов RDW (оценка

снизу)

Мы извлекли из базы данных «IP-страна» [9] диапазоны российских адресов. Суммарно, на момент июня 2005г., российская часть Интернета была представлена ~10 500 000 IP-адресов. Мы осуществили случайную выборку объемом в 105 000 уникальных адресов (1% от общего количества IP-адресов). Затем, используя сетевой сканнер nmap [14], машины, соответствующие адресам из выборки, были проверены на потенциальное наличие веб-сервера (мы проверяли открыт ли порт 80). У 1 379 машин порт 80 оказался открытым. Далее, мы пытались установить HTTP-соединение[12] с первыми 700[13] машинами. В случаях успешного соединения найденные сайты скачивались (с глубиной 3) и анализировались на наличие веб-интерфейсов. В результате мы идентифицировали **17** глубинных сайтов[14], каждый из которых предоставлял доступ к одной онлайн базе данных. Затем, получили оценку числа глубинных сайтов во всем Рунете:

$$\frac{1379}{700} \times 17 \times \frac{100}{1} \approx \mathbf{3350}$$

Метод случайной выборки из списка IP-адресов дает заниженную оценку потому, что не принимается во внимание виртуальный хостинг. В нашем случае это значит, что в ходе эксперимента мы пропустили определенное количество веб-сайтов, которые также расположены на 700 рассматриваемых машинах. Очевидно, какие-то из таких «пропущенных» сайтов могли быть глубинными. Принимая все это во внимание, можно утверждать, что ***RDW представлен не менее, чем 3350 ресурсами***.

5.2. Определение количества ресурсов RDW (выборка из хостов)

Для формирования репрезентативного списка хостов был использован набор данных «Хостграф»[15] от Яндекса. Из «Хостграфа» были извлечены все сайты, проиндексированные Яндексом. Также, для каждого извлеченного сайта был посчитан «индекс цитируемости», количество хостов, имеющих ссылки на данный сайт. Для повышения точности, из получившегося списка были удалены сайты, которые заведомо не являются глубинными. Например, мы удалили все сайты вида *.narod.ru, *.stih.ru и т.д.[16] Помимо этого, мы определили сайты в наиболее встречающихся доменах второго и третьего уровней. На основе просмотра и выборочной проверки этих сайтов мы произвели дальнейшую чистку нашего списка. В частности, мы удалили несколько групп хостов, ведущих к одним и тем же онлайн

базам данных. Например, были удалены все сайты вида *.mp3gate.ru (за исключением www.mp3gate.ru), так как любой из этих сайтов вел к базе данных на www.mp3gate.ru. В итоге, мы остановились на списке из 299 241 хоста. Вначале, мы решили проверить наше предположение о том, что среди высоко- и средне-цитируемых сайтов больше глубинных, чем среди мало-цитируемых сайтов. Для этого список был разбит на три части: 1) 49899 хостов с наибольшим цитированием (примерно одна шестая часть от списка); 2) 52103 хоста со средним цитированием (также примерно одна шестая часть); и 3) 197 239 хостов с наименьшим цитированием (все остальные хосты). В Таблице 1 приведены размеры случайных выборок и число найденных ресурсов для каждой из трех частей списка. 200 сайтов из нашей суммарной выборки скачивались (с глубиной 3) и анализировались на наличие веб-интерфейсов (сайты с веб-форумами не учитывались). Как мы видим, действительно, глубинные сайты попадаются чаще, если рассматриваются сайты с наибольшим цитированием.

Таблица 1

Список	1)	2)	3)
Размер выборки	50	50	100
Число найденных ресурсов	7	2	1

Далее, мы произвели более тщательный анализ, рассмотрев 288 сайтов. Размеры случайных выборок и число найденных ресурсов приведены в Таблице 2.

Таблица 2

Список	1)	2)	3)
Размер выборки	124	64	100
Число найденных ресурсов	14	3	0

Хосты, соответствующие глубинным сайтам, были рассмотрены более подробно. Наша задача заключалась в идентификации других хостов из списка, которые также ведут к найденным ресурсам. Для этого мы просматривали сайты вручную, а также использовали сервис [7], который возвращает список сайтов, расположенных на IP-адресе. В итоге, мы нашли, что, как минимум, у шести хостов из семнадцати есть аналоги в списке. Результаты приведены в Таблице 3. Аналоги выделены курсивом и помещены в соответствующую им часть списка. Также, в таблице указывается, сколько всего сайтов располагается на IP-адресе, который соответствует каждому рассмотренному хосту.

Таблица 3

	Число сайтов на IP-адресе	Список 1): 49899 хостов	Список 2): 52103 хоста	Список 3): 197239 хостов
www.boschbuy.ru	500	**X**		
www.avto-az.com	500	**X**		
www.1expo.ru	445	**X**		
www.all4auto.ru	60	**X**		*All4.ru*
www.stbcard.ru	-	**X** *web.stbcard.ru*		
www.komzdrav.ru	9	**X**		*Mosgorz-drav.ru*
www.oval.ru	1	**X**		
www.interproekt.ru	12	**X**		
melody.mton.ru	18	**X** *mton.ru*		*mtone.ru*
autoserver.ru	2	**X**		*audishop.ru*
www.mdo-tirus.ru	55	**X**		
63.ru	40	**X**		
www.ais-t.ru	499	**X**		
www.perspektiva.ru	11	**X**		*910777.ru*
www.pcpi.ru	5		**X**	
www.region34.ru	123		**X**	
www.mosfarm.ru	1		**X**	

Нетрудно получить оценки числа ресурсов для списков 1) и 2). Принимая во внимание то, что у двух глубинных сайтов из списка 1) есть «дубликаты», находящиеся в том же списке, получим оценку в 5230 и 2440 ресурсов для списка 1) и 2) соответственно. Помимо этого, к пяти глубинным ресурсам из списка 1) можно придти через сайты из списка 3) (см. Таблицу 3). Это значит, что примерно 5230*5/14 =~ 1860 ресурсов из списка 3) должно быть исключено, так как они ведут к уже посчитанным ресурсам из списка 1). В нашем случае, рассмотрев 100 случайных сайтов из списка 3) (см. Таблицу 2), мы не нашли ни одного ресурса. С учетом результатов предыдущей выборки (см. Таблицу 1), это значит, что, во-первых, объем выборки для списка 3) был недостаточен (желательно рассматривать не менее 200 сайтов), и, во-вторых, список 3) вероятно содержит около 2000 ресурсов (из которых большая часть является дубликатами ресурсов из списка 1)). Так или иначе, суммируя оценки списков 1) и 2) получаем общую оценку для всего Рунета: **7670 ресурсов**. Мы полагаем, что это число является оценкой сверху, и что погрешность определения не превысила 20%.

5.3. Оценка размеров для нескольких ресурсов RDW

Мы оценили размер одиннадцати онлайн баз данных (наиболее крупные в своих категориях из нам известных) в категориях «Авто», «Люди», «Музыка», «Справки» и «Товары и Услуги». Для определения среднего размера бд-наполненной страницы, мы рассматривали десять произвольных бд-наполненных страниц, содержащих информацию об одной записи или одной логической сущности[17]. Определяли размер каждой из десяти страниц, отбрасывали две страницы наибольшего размера и две наименьшего, затем, размеры оставшихся шести страниц усреднялись. Дополнительно, мы определяли какое количество бд-наполненных страниц проиндексировано Яндексом. Результаты приведены в Таблице 4.

Таблица 4

Глубинный сайт	Категория	Доступ	Число записей	Проиндексировано Яндексом	Средний размер страницы, КБ	Оценка размера базы данных, МБ
www.auto.ru	Авто	Своб.	86 900	<20	20,1	1 706
www.bibika.ru	Авто	Своб.	52 580	10 800	67,4	3 463
auto.vl.ru	Авто	Своб.	47 320	6 260	9,3	430
www.japancar.ru	Авто	Своб.	567 610	1 440	15,0	8 334
www.price.ru	Товары и Услуги	Своб.	4400000 [8700]	0	388,1	3 300
www.domoteka.ru	Товары и Услуги	Своб.	320000 [790]	0	363,0	280
pharm.mos.ru	Товары и Услуги	Своб.	891 920	15	25,4	22 101
www.mp3search.ru	Музыка	Своб.	[20 680]	13 300	50,3	1 016
pesenki.ru	Музыка	Своб.	150 000	960	29,5	4 316
encycl.yandex.ru /Бол.Сов.Энцикл./	Справки	Своб.	94541	-	7,5	697
www.pobediteli.ru	Люди	Своб.	1 008 740	0	7,2	7 129

Как мы видим, суммарное число записей в приведенных базах данных – около 7,5 миллионов (и суммарный размер более чем 50ГБ). Предполагая, что в RDW есть не менее 500 баз данных, сравнимых по числу записей с указанными в таблице[18], оценим общее количество записей в RDW. Результат - ~340 миллионов записей, что сопоставимо[19] с числом документов, проиндексированных Яндексом.
Завершая данный раздел, заметим, что самыми крупными ресурсами RDW (из нам известных) являются две библиотечные базы данных, доступные через поисковые интерфейсы на сайтах eLibrary.ru и Sigla.ru соответственно.

6. Обсуждение результатов

Проведенные эксперименты показали, что RDW представлен не более чем 7700 и не менее 3300 ресурсами. Это весьма обширный диапазон, но даже, исходя из него, можно утверждать, что число глубинных сайтов в Рунете на данный момент порядка 10^3 и не превышает 10 000.

В ходе работы мы несколько раз меняли наше отношение к методу случайной выборки из списка IP-адресов. Несмотря на недостатки, метод весьма прост и ясен. Он гораздо менее требователен в отношении исходных данных, чем метод случайной выборки из списка хостов. В последнем, получение и последующее формирование репрезентативного списка хостов может быть нелегкой задачей. Проблема виртуального хостинга, возникающая при использовании метода выборки из IP-адресов, может быть решена с помощью сервиса [7]. Данный сервис возвращает «все» домены, расположенные на указанном IP-адресе. Правда, нужно заметить, что проверка всех виртуальных хостов повышает трудоемкость метода. Например, как показано в Таблице 3, на семнадцати IP-адресах расположено более чем 2100 сайтов[20], каждый из которых необходимо рассмотреть.

Другой, более простой подход для внесения корректировки в оценку метода выборки из IP-адресов может заключаться в следующем: рассматриваем список глубинных сайтов; определяем IP-адреса машин на которых расположены сайты из списка; наконец, для каждого адреса проверяем, что возвращается в ответ на URL, содержащий IP-адрес (тот же глубинный сайт, другой сайт, страница по умолчанию веб-сервера, ошибка и т.д.). В работе мы осуществили такую проверку. Для этого использовался составленный нами список русскоязычных ресурсов. Мы получили, что только к ~55% ресурсам из этого списка можно придти по URL-ам с IP-адресами. Остальные 45% будут пропущены из-за того, что, в большинстве случаев, URL с IP-адресом возвращает не искомый глубинный сайт, а ошибку или ведет к другому сайту. Принимая во внимание то, что рассмотренный список ресурсов небольшой и формировался на неслучайной основе, мы не можем прямо использовать полученную оценку. Однако, полагаем, что метод случайной выборки из списка IP-адресов учитывает от 45% до 65% всех глубинных сайтов (и, соответственно, игнорирует от 35 до 55% ресурсов). Применив данную корректировку к нашей оценке из раздела 5.1, получим следующий диапазон: 3350*100/65 – 3350*100/45. Заметим, что верхняя граница диапазона не превышает оценки, полученной в разделе 5.2. Этот же диапазон и укажем как нашу заключительную оценку количества

глубинных сайтов в RDW - **RDW включает в себя 5100-7500 ресурсов**.

Для сравнения количества ресурсов в DW и RDW рассмотрим результаты, полученные в работе [5] (около 300 000 глубинных сайтов и 450 000 онлайн баз данных в DW (на момент апреля 2004г.)), и нашу оценку из раздела 5.1 (3350 глубинных сайтов в RDW (на момент июня 2005г.)). В силу того, что использовался один и тот же метод, эти результаты можно сравнить[21]. Нетрудно увидеть, что с учетом роста числа ресурсов DW с апреля 2004г. по июнь 2005г., RDW примерно сотая часть от DW. Что же касается, размера онлайн баз данных в DW и RDW, то мы полагаем, что средняя онлайн база данных в DW содержит больше записей, чем средняя база данных в RDW. Это может быть объяснено рядом причин. Во-первых, базы данных, доступные онлайн, стали появляться в Рунете в целом позже, и, следовательно, имели меньше времени для наполнения данными. Другая причина – различия в уровне и образе жизни усредненного веб-пользователя. Действительно, значительная часть баз данных в DW являются «потребительскими», т.е. предоставляющими информацию о товарах и услугах, автомобилях, недвижимости и т.д. Таким образом, такой показатель как, скажем, количество автомобилей на душу населения естественно влияет на количество записей в базах данных, посвященных автомобилям. Для наглядной демонстрации, сравним базу данных на Auto.ru и базу данных на Autotrader.com (поиск по автомобилям в США/Канаде) – они содержат менее 90 тысяч и около 2,5 миллионов записей соответственно.

7. Заключение

В данной работе мы изучали русскоязычную часть глубинного Веба. Нашей основной целью была количественная оценка числа глубинных сайтов в RDW. Для получения оценки использовалось два метода: метод случайной выборки из списка IP-адресов и, предложенный нами, метод случайной выборки из списка хостов. Важную роль в методе выборки из хостов сыграл репрезентативный список хостов, полученный нами с помощью набора данных «Хостграф». Согласно полученным оценкам и их корректировкам, RDW представлен 5100-7500 глубинными сайтами. Мы также показали, что среди высоко-цитируемых сайтов значительно больше глубинных сайтов, чем среди мало-цитируемых. Сравнение с данными из работы [5] показало, что доля RDW во всем глубинном Вебе составляет около одного процента.

Мы надеемся, что данное исследование будет интересно многим

русскоязычным исследователям и разработчикам, работающим в области извлечения веб-данных, а полученные характеристики DW будут являться стимулом для дальнейших исследований, посвященных DW и работе с данными из DW.

8. Литература

1. R. Baumgartner, S. Flesca, and G. Gottlob. *Visual Web Information Extraction with Lixto*. In Proc. of 27th Int.Conf. on Very Large Data Bases (VLDB'01), 2001
2. M.K. Bergman. *The Deep Web: Surfacing Hidden Value*. Journal of Electronic Publishing, 7(1), 2001
3. *CompletePlanet: The Deep Web Directory*. http://completeplanet.com
4. V. Crescenzi, G. Mecca, and P. Merialdo. *RoadRunner: Towards Automatic Data Extraction from Large Web Sites*. VLDB Journal, p.109-18, 2001
5. K.C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. *Structured Databases on the Web: Observations and Implications*. SIGMOD Record, 33(3), 2004
6. H. Davulcu, S. Koduri, and S. Nagarajan. *DataRover: A Taxonomy Based Crawler for Automated Data Extraction from Data-Intensive Websites*. In Fifth International Workshop on Web Information and Data Management (WIDM'03), 2003
7. *DomainsDB.net: Reverse IP Lookup, Reverse NS Lookup, Whois Tools*. http://domainsdb.net
8. D. Florescu, A.Y. Levy, and A.O. Mendelzon. *Database Techniques for the World-Wide Web: A Survey*. SIGMOD Record, 27(3), 1998
9. *GeoIP Free Country database*. http://www.maxmind.com/app/geoip_country
10. *Invisible Web*. http://invisibleweb.com
11. O. Kaljuvee, O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. *Efficient Web Form Entry on PDAs*. In Proc. of 10th Int. WWW Conf., 2001
12. A.H.F. Laender, B. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira. *A Brief Survey of Web Data Extraction Tools*. SIGMOD Record, 31(2), 2002
13. S. Lawrence and C. Giles. *Searching the World-Wide Web*. Science, 280(4), p.98-100, 1998
14. *Nmap: Network Security Scanner*. http://www.insecure.org/nmap/
15. *Search Web by Domain*. http://searchdns.netcraft.com/?host
16. S. Raghavan and H. Garcia-Molina. *Crawling the Hidden Web*. In

Proc. of 27th Int.Conf. on Very Large Data Bases (VLDB'01), 2001
17. C. Sherman and G. Price. *Invisible Web: Uncovering Information Sources Search Engines Can't See*. ISBN: 091096551X, 2001
18. D. Shestakov, S.S. Bhowmick, and E.-P. Lim. *DEQUE: Querying the Deep Web*. Data and Knowledge Engineering Journal (DKE), 52(3), 2005

Characterization of Russian Deep Web

Denis Shestakov[1], Natalia Vorontsova[2]

[1] University of Turku, Turku, Finland, da_shestakov@yahoo.com
[2] IM Sechenov Moscow Medical Academy, Moscow, Russia, n_vorontsova@myrealbox.com

The significant portion of the Web is hidden behind search forms and not indexed by conventional search engines. This part of the Web is known as the deep Web. Pages in the deep Web are dynamically generated in response to queries submitted via search forms. In this work, we studied the Russian part of deep Web. Our main goal was to estimate the number of deep Web sites in the Russian deep Web. The presented study is a first work devoted to the certain part of deep Web, which is formed on the basis of some particular language usage.

[1] Т.е. сгенерированными веб-сервером после соответствующего запроса и возвращаемыми пользователю/приложению как результат этого запроса.
[2] Очевидно, такое отождествление неверно в общем случае, так как HTML форма лишь наиболее распространенный способ организации веб-интерфейса к базе данных.
[3] Отсюда и два других названия DW: скрытый и невидимый Веб.
[4] Мы полагаем, что базы данных, чьи веб-интерфейсы неизвестны поисковой системе, предназначены для строго ограниченного числа пользователей (например, внутри интранет-сети какой-нибудь компании) и, поэтому, не являются частью Веба.
[5] Определение числа онлайн баз данных по числу различных веб-интерфейсов на глубинном сайте тоже неоднозначно. Например, на Auto.ru расположено более 100 веб-интерфейсов, которые позволяют искать исключительно среди подержанных/новых автомобилей определенной марки. Поэтому, если считать, что существует база данных исключительно по подержанным автомобилям, то можно говорить и о существовании серии баз данных, каждая из которых посвящена подержанным автомобилям только какой-то одной марки.
[6] Мы придерживаемся именно такой позиции.
[7] Полученная оценка является весьма приблизительной, т.к. для каждой рассмотренной онлайн базы данных было проверено всего несколько динамических страниц.
[8] Данный список был не полон (не было учтено около 200 миллионов IP-адресов), т.к. Chang et al. использовали список действительных IP-адресов на момент 1997г. Тем не менее, данная ошибка является технической, причем, исходя из особенностей метода, лишь незначительно влияет на итоговые результаты.
[9] Существует поиск от Netcraft [15], позволяющий просматривать около 65 миллионов (на момент июня) доменных имен, но требуются специальные ухищрения, чтобы получить с его помощью репрезентативный список, т.к. при каждом поиске возвращается не более 500 результатов. Также, в случае Рунета, источник на http://stylehost.ru/index.php?ct=ru содержит около 100 тысяч доменов второго уровня в зоне ru.
[10] Или к одним и тем же базам данных, если считать, что на Auto.ru расположено несколько баз данных.
[11] Рассматриваются только файлы определенных форматов, а именно форматов, известных поисковым системам. Среди наиболее часто встречающихся форматы doc, txt, pdf, ppt, xls. Аудио/видео, графические и архивные файлы не учитываются, т.к. их содержимое не индексируется сегодняшними поисковыми системами.
[12] В качестве URL использовался IP-адрес. В случае ошибки использовали не пустое значение, возвращаемое функцией *gethostbyaddr*.
[13] Т.е. чуть больше половины машин с открытым портом 80. Остальная половина не была рассмотрена из-за недостатка времени..
[14] А также, считавшиеся отдельно, 21 сайт с веб-форумами.
[15] Хостграф - граф ссылок между хостами, известных поисковой системе Яндекс. Все ссылки, исходящие с разных страниц одного хоста на разные страницы другого хоста, заменяются на одну.
[16] В основном это сайты на различных бесплатных хостингах.

[17] Музыкальные альбомы для баз данных в категории «Музыка» и прайс-листы компаний для баз в категории «Товары и Услуги». Количество сущностей указано в квадратных скобках.
[18] Что, вполне вероятно, не так. В частности, приведенные базы данных являются одними из самых крупных в своих категориях. Впрочем, в RDW есть и очень крупные ресурсы – например, eLibrary.ru, предоставляет ограниченный доступ к ~8 000 000 документов с суммарным размером около 1 ТБ.
[19] Более чем в два раза меньше.
[20] Часть из которых, возможно, не является частью Рунета.
[21] Мы используем именно нескорректированную оценку, что была получена в разделе 5.1, так как в работе [5] коррекция не производилась.