Бизнес-аналитик часто сталкивается с ситуацией, когда ему известно о существовании в веб-пространстве какого-то документа, но не может найти его с помощью традиционных поисковиков, какими сегодня можно считать такие системы, как Google, Yahoo, Bing, Яндекс, Рамблер или Мета. Однако, вспомнив или найдя в закладках адрес (URL) этого документа, он без труда выходит на него. То есть в веб-пространстве этот документ есть, а найти его привычным способом нельзя. Пользователь столкнулся с невидимым (invisible) для поисковых систем ресурсом

Глубинный web – информационная среда для бизнес-аналитика

Что такое глубинный веб?

Подобные ресурсы уже давно имеют собственное название – «глубинный web» – которое ввел Джилл Иллсворт (Jill Ellsworth) в 1994 году, обозначив им источники, недоступные пользователям обычных поисковых систем.

Под термином «глубинным web» (invisible web, deep web, hidden web) принято понимать часть web-пространства, не индексируемую роботами (web crawlers) поисковых систем. Именно таким документом из глубинного web оказался ресурс, необходимый пользователю-аналитику. Используя аналогию, информация, будучи недоступной для поиска, находится «в глубине» (англ. – *deep*). При этом не стоит путать глубинный web с ресурсами, вовсе недоступными через Интернет – это темный web (dark web), и речь о нем здесь идти не будет. Не будем обсуждать и ресурсы, доступ к которым открыт лишь для зарегистрированных пользователей, хотя такие ресурсы также относятся к глубинному web.

В 2000 году американская компания BrightPlanet (www.brightplanet.com) опубликовала сенсационный доклад, в котором утверждается, что в web-пространстве в сотни раз больше страниц, чем их удалось проиндексировать самыми популярными на то время поисковыми системами. Компания разработала программу LexiBot, которая позволяет сканировать некоторые динамические web-станицы, формируемые из баз данных, и, запустив ее, получила неожиданные данные. Выяснилось, что в глубинном web находится в 500 раз больше документов, чем доступно через поисковые системы. Конечно, эти цифры неточны. Кроме того, стало известно, что средняя страница глубинного web на 27% компактней средней страницы из видимой части web-пространства.

Сегодня ситуация изменилась, например, ведущие поисковые системы могут индексировать документы, представленные в форматах, содержащих текст. Конечно, это прежде всего pdf, rtf и doc. В 2006 году Google запатентовали способ поиска в глубинном web: «Searching through content which is accessible through web-based forms» (рис. 1). Сейчас, по мнению разных авторов, к видимому web уже относится порядка 20-30% содержимого web-пространства, а это означает, что пользователям традиционных поисковых систем оказывается доступным еще больший объем информации, размещенной в Сети.

Причины возникновения

В глубинном web находятся web-ресурсы, не связанные с остальными гиперссылками – например, страницы, динамически создаваемые по запросам к базам данных, документы из баз данных, доступные пользователям через поисковые web-формы (но не по гиперссылкам). Такие документы остаются недоступными для робота, неспособного в режиме реального времени правильно заполнить поля формы значениями (формировать запросы к базам данных).

Вот что говорят о глубинном web Крис Шерман и Гэри Прайс в своей книге «The Invisible Web: Uncovering Information Sources Search Engines Can't See»:

«Большинство страниц невидимого Интернета могут быть проиндексированы технически, но не индексируются, потому что поисковые системы решили их не индексировать… Большинство «невидимых» сайтов имеют высококачественный контент. Просто эти ресурсы не могут быть найдены с помощью поисковых машин общего назначения…

…Некоторые сайты используют технологию баз данных, что действительно сложно для поисковой машины. Другие сайты, однако, используют сочетание файлов, которые содержат текст и мультимедиа, а поэтому часть из них может быть проиндексирована, а часть – нет.

…Некоторые сайты могут быть проиндексированы поисковыми машинами, но это не делается потому, что поисковые машины считают это непрактичным – например, по причине стоимости или потому, что данные настолько короткоживущие, что индексировать их просто бессмысленно – например, прогноз погоды, точное время прибытия конкретного самолета, совершившего посадку в аэропорту и т. п.»

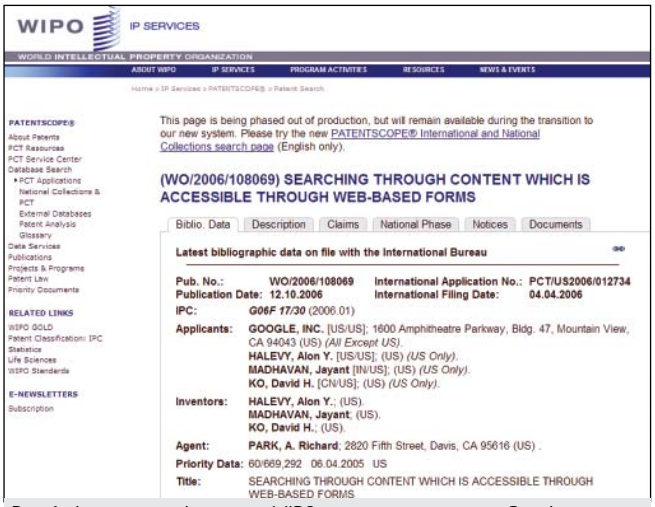

Рис. 1. Фрагмент web-ресурса WIPO с описанием патента Google на поиск в глубинном web

Основные ограничения, связанные с роботами поисковых машин, можно объяснить следующими основными причинами: для публичных поисковых служб важнее обеспечить точность поиска, чем полноту, важнее обеспечить получение ответа на запрос в приемлемое время, чем точность. Отсюда – ограничения на глубину сканирования web-ресурсов, попытки «фильтрации» контента по содержанию, отсеивание страниц, содержащих излишние выходные гиперссылки и т. п. При этом часто с водой выплескивается и ребенок. Общепризнано, что ценность ресурсов глубинного web зачастую выше ценности ресурсов видимой части web-пространства.

Можно упомянуть еще один источник пополнения глубинного web – владельцы сознательно не хотят, чтобы их web-ресурсы находили с помощью поисковых систем. Чаще всего такие web-ресурсы представляют нечто не совсем законное, хакерские форумы, архивы неавторизованного контента и т. п. Понятно, что многие из таких ресурсов очень интересны для изучения бизнес-аналитикам.

Многие компании сначала подключаются к общей Сети, и лишь потом тратят большие средства на защиту. Владельцы сайтов могут попытаться запретить индексацию тех или иных страниц своих ресурсов, прописав запрещающую команду в файле robots. txt, но поисковые системы могут ее проигнорировать. Поэтому такие ресурсы удаляют, либо удаляют гиперссылки, переводя ресурсы в глубинный web. Например, недавно бизнес-каталоги Auto.ru и Drom.ru отказались отдавать свои объявления «Яндексу», то есть, защищая свои информационные активы, компании перевели свои ресурсы в глубинный web.

Виды ресурсов глубинного web

Существует несколько типов ресурсов – глубинного web, например, как было замечено, это могут быть быстро устаревающие web-страницы. Кроме того, к глубинному web относятся web-ресурсы, представляющие собой мультимедиа информацию. Как известно, в данное время еще не существует удовлетворительных алгоритмов поис-

ка не текстовой информации. Динамически генерируемые по запросу страницы также часто попадают в глубинный web. Зачастую таких страниц без запроса не существует, они генерируются при запросе к базам данных. Получается, что информация, вроде бы в web-пространстве и имеется, но возникает она лишь в момент обработки запроса, а универсального алгоритма заполнения роботами поисковых форм пока не существует. И, наконец, если на web-ресурс не ведут никакие ссылки, то роботы поисковых систем никаким образом не могут узнать о его существовании.

Основатель BrightPlanet Майкл Бергмана (Michael K. Bergman) смог выделить 12 разновидностей глубинных web-ресурсов, относящихся к классу онлайновых баз данных. В списке оказались как традиционные базы данных (патенты, медицина и финансы), так и публичные ресурсы – объявления о поиске работы, чаты, библиотеки, справочники. Бергман причислил к глубинным ресурсам и специализированные поисковые системы, которые обслуживают определенные отрасли или рынки, базы данных которых не включаются в глобальные каталоги традиционных поисковых служб.

К глубинному web также относятся многочисленные системы интерактивного взаимодействия с пользователями – помощи, консультирования, обучения, требующие участия людей для формирования динамических ответов от серверов. К ним также можно отнести и закрытую (полностью или частично) информацию, доступную, пользователям Сети только с определенных адресов, групп адресов, иногда городов или стран. К «скрытой» части Сети многие причисляют и web-страницы, зарегистрированные на бесплатных серверах, которые индексируются, в лучшем случае, лишь частично – поисковые системы во избежание рекламного спама не стремятся обходить их в полном объеме.

К глубинному web также относится категория так называемых «серых» сайтов, функционирующих на основе динамических систем управления контентом (Dynamic Content Management Systems). В поисковых системах обычно ограничивается глубина индексирования таких сайтов во избежание возможного циклического просмотра одних и тех же страниц.

Примеры ресурсов глубинного web

Как же найти web-ресурсы, размещенные в глубинном web? Если ресурсы требуют заполнения специальных форм, дополненных, например, капчами (captcha – нечетким графическим изображением букв и цифр, которые требуется ввести с клавиатуры в определенное поле), то необходимо выйти на базу данных, предположительно содержащую необходимые документы. Найти базы данных – источники скрытого web – можно с помощью обычных поисковых систем, обобщив запрос и введя уточняющие слова, такие как «база данных», «банк данных», database и т. п.

Приведем общеизвестный пример: пользователю требуется статистика по катастрофам самолетов в Аргентине. Естественный запрос к традиционной поисковой системе выдает огромный список газетных заголовков, а на запрос «aviation database», можно сразу выйти на базу данных NTSB Aviation Accident Database (http://www.ntsb.gov/ntsb/query. asp).

Для поиска в «скрытой» Сети, а именно в том ее сегменте, который составляют базы данных, сегодня уже существуют некоторые специализированные ресурсы. Лидером среди навигаторов в «скрытом» web является сайт CompletePlanet (www.completeplanet.com) компании BrightPlanet. Этот сайт является крупнейшим каталогом, насчитывающим свыше 100 тысяч ссылок. Компания BrightPlanet также создала персональную утилиту для поиска в онлайновых базах данных – LexiBot, которая может обеспечивать поиск в нескольких тысячах поисковых систем «скрытого» web. Метопоисковый пакет DeepQueryManager (DQM) этой же компании обеспечивает поиск более чем по 70 тысячам «скрытым» web-ресурсам.

Исследование, проведенное еще в 2006 году (B. He, M. Patel, Z. Zhang, K. Chang, CACM 2006), показало, что глубинный web охватывает более 300 тыс. сайтов, связанных с более чем 450 тыс. баз данных, не охватываемых традиционными поисковыми системами. К наиболее интересным для бизнес-аналитиков ресурсам глубинного web относятся: базы данных на юридические и физические лица; отраслевые базы данных; репутационные базы данных (черные и белые списки); криминалогические базы данных; базы данных товаров и услуг; каталоги продукции и т. п. К всемирно известным бизнес-ресурсам, размещенным в глубинном веб, относятся: *amazon.com, ebay.com, realtor.com, cars.com, imdb.com.*

Приведем еще несколько примеров баз данных и каталогов глубинного web:

- PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) – с ресурса обеспечивается доступ к свыше 14 млн. ссылок системы MEDLINE, включая ссылки на полные тексты статей и информационные ресурсы. Имеется возможность перехода к службе PubMed Central (PMC), к свободно доступному архиву статей (свыше 90 тысяч) из научных журналов. Обеспечивается также доступ к глобальной поисковой системе NCBI, охватывающей базы данных по естествознанию.
- LookSmart's FindArticles (http://www.findarticles.com) – база данных FindArticles – архив, содержащий 2,8 млн. статей из более 500 источников, накапливаемый с 1998 года.
- Librarians' Index to the Internet (http://lii.org) – каталог, содержащий свыше 14000 Интернет-ресурсов и включающий ссылки на «скрытые» в web-пространстве базы данных. У владельцев таких баз данных есть возможность поместить соответствующую гиперссылку в этом каталоге на свой ресурс (в LII есть ссылка «and databases» (добавить базу данных).
- FindLaw (http://www.findlaw.com) – один из наиболее популярных в мире юридических web-сайтов – огромный каталог правовых ресурсов, содержащий аннотирован-

ный список свободно доступных баз данных нормативно-правовых документов, для которых данный ресурс является «точкой входа».

- About.com (http://www.about.com) – портал, охватывающий тысячи снабженных комментариями ссылок на web-ресурсы, в том числе и на ресурсы глубинного web (имеется ссылка «Invisible Web»). На портале предоставляется возможность поиска в каталоге. Ресурс также включает несколько статей по проблематике глубинного web: «What is the Invisible Web?», «Finding the Invisible Web», «Top Places to Search the Invisible Web» и др.

Особенность большинства «глубинных» ресурсов – в их узкой специализации. Для поиска в них используются те же механизмы, что и для «поверхностного» web, однако, чаще всего, роботы поисковых систем для глубинного web включают уникальные для каждого такого ресурса модули доступа к данным.

Традиционная поисковая система чаще всего может выдать адрес базы данных, но не скажет, какие документы конкретно содержаться в ней. Типичный пример – информационно-поисковые системы по украинскому (http://zakon.rada.gov.ua) или российскому законодательству (http://www.kodeks.ru). Тысячи документов из баз данных становятся доступны только после входа в систему, а роботы стандартных поисковых систем не в состоянии заиндексировать контент баз данных.

Парадоксально, но как один из ресурсов глубинного web можно рассматривать и архив ресурсов открытого web-пространства. Такой архив – «Internet Archive» с 1996 года создает компания Alexa (www.archive.org). Сегодня объем базы данных Alexa превышает 150 млрд. web-страниц. Технология хранилища Alexa включает ряд современных средств управления гигантским документальным хранилищем. Например, с помощью технологии Alexa выполняется кластеризация web-ресурсов, т.е. формирование коллекций документов, близких по тематикам. Особый интерес у пользователей сервиса Alexa вызывает «Машина времени» (Wayback Machine), открывающая доступ к временным срезам web-пространства. Одно из наиболее интересных практических применений этой технологии – восстановление документов, некогда опубликованных в web-пространстве, но впоследствии удаленных. При этом рост глубинного web грозит серьезными пробелами в хранилище системы, связанными с увеличивающимся количеством сайтов, эксплуатирующих различные технологии управления контентом, динамической публикацией документов из баз данных и т. п.

Сервисы работы с глубинным web

Традиционные поисковые системы стремятся (в меру своих возможностей) сузить пространство глубинного web, постепенно захватывая такие ниши, как блоги, научные сайты, информационные агентства. Так в качестве

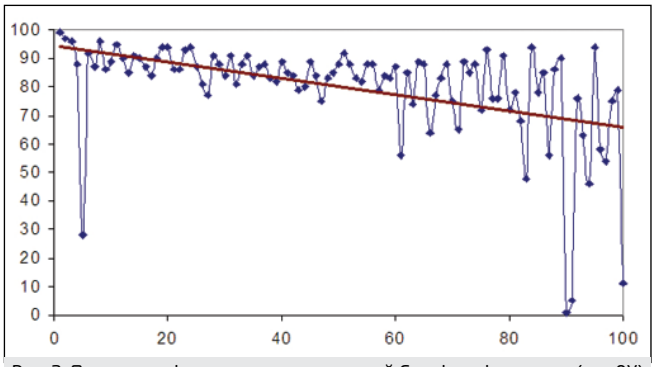

Рис. 2. Процент зафильтрованных системой Google web-страниц (ось OY) для различных сайтов, ранжированных по размеру в порядке убывания. Ось OX – номер сайта в ранжированном списке

вспомогательных сервисов для поиска по глубинному веб от Google можно рекомендовать: Google Book Search (поиск книг), Google Scholar (поиск научных публикаций), Google Code Search (поиск программного кода).

Система Goldfire Research (Invention Machine Corp.) позволяет обрабатывать контент глубинного web, размещенный на более чем 2000 сайтов правительственных, академических, исследовательских и коммерческих организаций США. Goldfire Research обладает информацией о механизмах доступа к базам данных глубинного web и автоматически генерирует запросы к ним.

Исследовательская поисковая система Infovell (University of California at Berkeley) позволяет искать в глубинном web по «ключевым фразами» от параграфов до целых документов или даже наборов документов общим объемом до 25 тысяч слов. Infovell не зависит от языка, пользователи могут искать страницы на английском, арабском, китайском языках или же вводить в строке поиска математические уравнения, химические формулы.

Российская компания «Р-Техно» создала систему «it2b. интернетошпионопаук 3000+», предназначенную для выгрузки данных из невидимого сегмента Интернета. На основе этой системы построен поисковый сервис Web Insight (http://r-techno.com/rtechno/online-services/webinsight), обеспечивающий поиск по официальным сайтам и базам данных России и ближнего зарубежья, а именно, по документам ФНС, ФССП, Пенсионного фонда, ФАС, Трудовой инспекции, ФРС, МЧС, Арбитражного суда, МВД, ФСБ. Известны такие базы данных «Р-Техно», как «Розыск Интерпола»; «Компании США уличенные в мошенничестве»; «Недобросовестные поставщики ФАС», «Должники металлургической отрасли» и т. п.

Существующие средства анализа и продвижения web-ресурсов позволяют по-новому подойти к оценке соотношения объемов видимого и глубинного web. Так на сайте http://www.cy-pr.com/(не являющемся поисковой системой) приводится информация о реальном количестве документов на исследуемом веб-сайте, представленном в RUNet, и о количестве документов, заиндексированных различными поисковыми системами, в том числе Google и Яндекс. Получив репрезентабельную выборку по сайтам, например, по рейтингу Рамблера top100 (http://top100.rambler.ru), можно получить оценку соотношения видимой и глубинной части в RUNet-сегменте web-пространства.

На рис. 2 приведен график зависимости доли web-страниц, попавших в глубинную часть web (зафильтрованных поисковой системой Google страниц), к общему размеру веб-ресурса для 100 сайтов из категории «Новости и СМИ» рейтинга Рамблера top100, таких как РосБизнесКонсалтинг, Лента. ру, РИА «Новости», Газета. ру и т. д. На графике сайты ранжировались по объему – общему количеству страниц.

Как показывают расчеты, объем информации, оказавшейся в глубинной части web-пространства, превышает объем информации из видимой части примерно в 7 раз. Оказывается, за редким исключением, чем крупнее ресурс, тем большая его часть относится к глубинному веб. В этом смысле небольшие web-ресурсы выигрывают в доступности. Так как большая доля новостных документов оказывается в глубинном web, для задач бизнес-аналитики требуется специальный сервис доступа к такой информации. Именно такой сервис предоставляют службы интеграции новостного контента – архивы сетевых СМИ. Российские и украинские бизнес-аналитики активно используют крупнейшие архивы информации из открытых источников «Интегрум» (http://integrum.ru) и InfoStream (http://infostream.ua). Именно использование открытых источников позволяет Конкурентной разведке действовать в рамках правового поля, и при этом иметь высокую эффективность. Известно, что американские разведслужбы получают из открытых источников до 95% всех разведданных при затратах на OSINT (Open Source INTelligence) около 1% из всего бюджета на разведку.

Можно констатировать, что чем быстрее растет web-пространство, тем хуже оно охватывается традиционными каталогами и поисковым машинам. Ввиду роста количества web-сайтов и порталов (использующих базы данных), динамических систем управления контентом, появления новых версий форматов представления информации, глубинный сегмент web растет очень интенсивно. С одной стороны, Интернет как огромное хранилище увеличивает объем информации, доступной «в принципе», но с другой стороны – растет информационный хаос, увеличивается энтропия сетевого информационного пространства. Все меньшая часть информационных ресурсов становится доступной пользователям реально.

Ведущие поисковые системы по-прежнему пытаются найти технические возможности для индексации содержимого баз данных и доступа к закрытым web-сайтам, однако, их задачи объективно расходятся с задачами бизнес-аналитиков – ориентация традиционных поисковых служб на массовый сервис в данном случае оправдана. Таким образом, ниша для систем поиска в глубинном web становится все шире.

Дмитрий Ландэ,
заместитель директора Информационного центра «ЭЛВИСТИ»