УДК 004.77

Д. В. Ландэ, Б. А. Березин, В. А. Додонов
Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина

Обзор особенностей и возможности контент-мониторинга национального сегмента сети Интернет

На примере китайского сегмента сетевого информационного пространства рассмотрены основные сравнительные характеристики представления национальных информационных ресурсов и мирового сегмента Интернета. Показаны возможности сбора контента китайского сегмента с использованием каналов RSS и программных средств системы мониторинга веб-ресурсов.

Ключевые слова: *интернет-контент китайского сегмента Интернета, мировой сегмент Интернета, каналы RSS, система мониторинга веб-ресурсов.*

Актуальность и постановка проблемы

В настоящее время китайский сегмент Интернета является наибольшим в мире по количеству пользователей — более 688 млн (что составляет больше 50 % населения страны) и быстрорастущим сегментом Сети. Третий по количеству пользователей (после Китая и Индии) сегмент Интернета США насчитывает около 280 млн пользователей, что составляет более 80 % населения страны. В ряде работ [1–3] отмечаются особенности китайского сегмента Интернета: большое число мобильных интернет-пользователей — в Китае они составляют около 90 % владельцев смартфонов, а в США около 40 %; большая активность и стабильность в публикации контента Интернета (для пользователей из группы стран, включающей Китай, среднее число публикаций на 20–50 % больше группы стран, включающей США); возраст основных групп пользователей — около 30 % составляет 20–29 лет, около 22 % составляет 10–19 лет и около 24 % составляет 30–39 лет. Распределение доли пользователей мировой сети Интернет по странам показано на рис. 1 (приведены первые 37 стран из около двухсот по убыванию доли пользователей, использованы данные [4]). На рисунке каждой стране соответствует заштрихованный столбец, отражающий количество пользователей Интернетом в стране. Не заштрихованная часть столбца отражает население страны, не использующее Интернет, т.е., из графика видно вклад пользователей Интернетом каждой страны в численность мировой Сети, а также степень использования Интернета внутри страны. Рост числа пользователей китайского сегмента Интернета отражен

на рис. 2 (на основе данных работы [5]). По оси абсцисс на рисунке представлены года, а по оси ординат — миллионы пользователей.

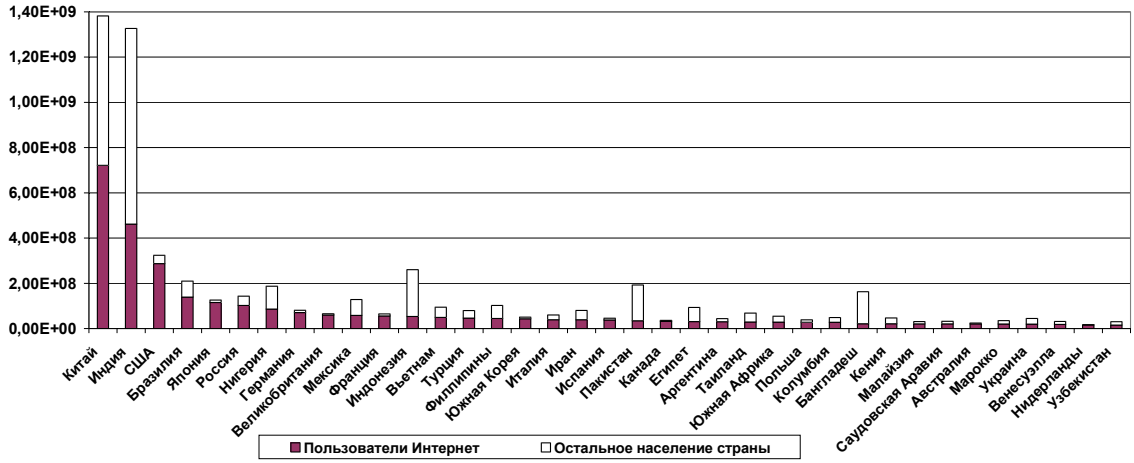

Рис. 1. Распределение доли пользователей мировой сети Интернет по странам

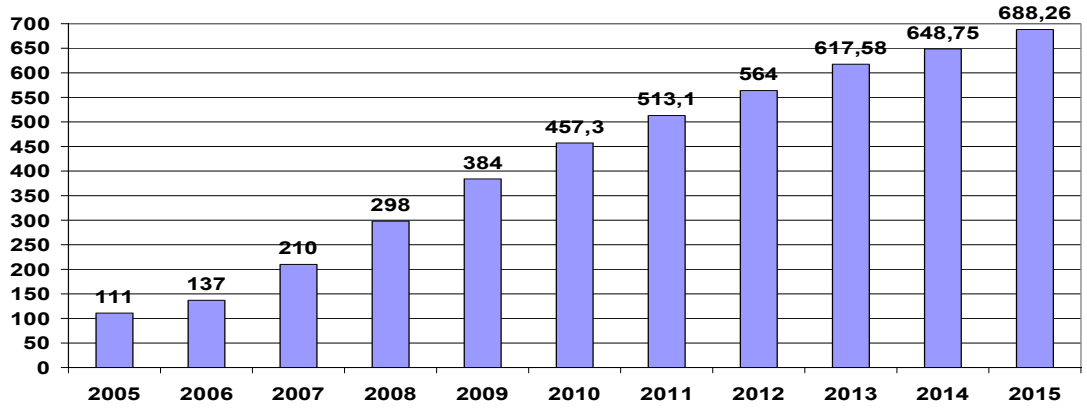

Рис. 2. Рост количества пользователей китайского сегмента Интернета по годам (в млн пользователей)

Контент китайского сегмента Интернета представлен 4,23 млн веб-сайтов и 212,3 млрд веб-страниц. Ежегодный рост их количества отражен на рис. 3 (на основе данных работы [5]). По оси абсцисс на рисунке представлены года, ось ординат слева (в млн) относится к колонкам роста числа сайтов, а ось ординат справа (в млрд) относится к кривой роста числа веб-страниц.

Преимущественное использование китайского языка и незначительная доля английского в контенте китайского сегмента Интернета затрудняет непосредственное использование китайского контента в европейских и американских странах. Однако возможности Google-переводчика позволяют преодолеть языковый барьер и обуславливают актуальность сбора контента китайского сегмента Интернета.

Приведенные выше особенности китайского сегмента Интернета делают перспективным сбор и использование его контента в различных направлениях. Однако, в настоящее время, особенности контента и возможности его сбора ис-

следованы недостаточно. Имеющиеся работы [1–5] и др., как правило, посвящены лишь отдельным характеристикам контента. В данной работе рассматриваются основные сравнительные характеристики веб-ресурсов китайского и мирового сегментов Интернета. Показываются возможности сбора контента китайского сегмента с использованием каналов RSS и программных средств системы мониторинга веб-ресурсов.

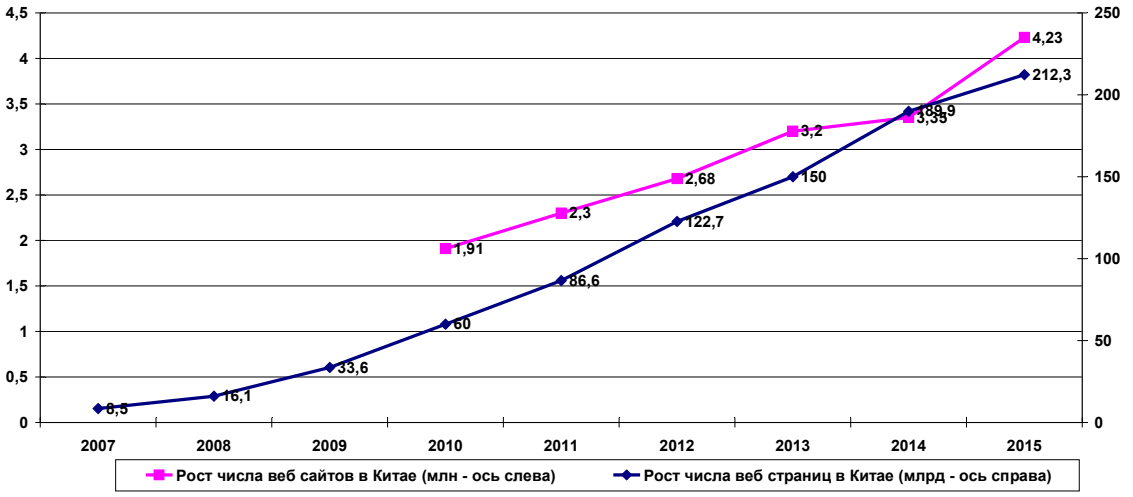

Рис. 3. Рост количества веб-сайтов (в миллионах) и веб-страниц (в миллиардах) в китайском сегменте Интернета по годам

Обзор особенностей контента

Особенности контента китайского сегмента Интернета и возможности его сбора определяются рядом характеристик, среди которых: количество веб-сайтов и веб-страниц, их распределение по регионам Китая; периодичность обновления веб-страниц; языки и кодировки, а также форматы данных, используемые при подготовке веб-документов и мультимедиа; основные порталы газет, информагенств, учебных и научных заведений с репозитариями открытых публикаций, социальные сети и т.д.

Вышеприведенные характеристики контента исследуются в ряде работ [1–5]. При анализе характеристик контента надо учитывать, что доступ к веб-сайтам часто происходит не непосредственно по URL-адресу, а через поисковые системы, т.е., зависит от индексации веб-сайтов в них. В работах [6, 7] исследуется покрытие разных поисковых систем при доступе к веб-контенту таких стран как США, Китай и Тайвань, Сингапур. Показано, что процент сайтов, к которым обеспечивается доступ через поисковые системы, зависит от конкретной поисковой системы, от страны, где размещаются сайты, а также от типа самих сайтов (коммерческие сайты, сайты органов власти, сайты организаций, университетов). В работах [8–11] рассматриваются вопросы видимости такого типа контента как институциональные репозитории и оценка их индексации в поисковых системах Google и Google Scholar с использованием операторов расширенного поиска и сравнением количества найденных документов. Аналогичные средства используют также при

проверке индексации сайтов в сервисах поисковой оптимизации SEO (search engine optimization). Учитывая результаты работ [6–11], при анализе особенностей контента китайского сегмента Интернета, кроме данных из работ [1–5] и [12–26], рассматриваются сравнительные оценки этих характеристик, полученные с помощью операторов поиска при доступе к контенту через разные поисковые системы.

Количество веб-сайтов. В [12] приводятся данные, что на конец 2010 г. численность веб-сайтов в Китае составляла 1,91 млн. По данным [5], в конце 2015 г. их количество составило 4,23 млн. На рис. 4 представлено распределение этого количества веб-сайтов по регионам Китая. По оси абсцисс показаны регионы в порядке убывания количества сайтов. По оси ординат — количество веб-сайтов. Из графика видно, что количество сайтов меняется от почти 671 тыс. для провинции Guangdong (15,9 % от общего количества сайтов, население 86 420 000), до примерно 1 тыс. для автономного района Tibet (население 2 740 000).

Сравнить количество веб-сайтов китайского контента с объемом мирового контента можно на основе данных [13–15]. В [13] приведено общее количество веб-сайтов в мире, составляющее более 1,08 млрд (середина сентября 2016 г.). Компания Netcraft начала анализировать рост числа веб-сайтов в Интернете с 1995 г., а с 2000 г., учитывая большое число сайтов, создаваемых автоматически при регистрации доменов и т.д., стала отдельно регистрировать общее количество сайтов и количество активных сайтов (https://www.netcraft.com/active-sites). В обзоре веб-серверов за август 2016 г. [14] приведено общее количество 1 153 659 413 сайтов и 17 3007 991 активных сайтов. Таким образом, 4,23 млн веб-сайтов китайского контента составляют около 2,4 % от числа активных сайтов в мире. По оценкам [15] в украинском сегменте Интернета в мае 2013 г. было около 532,7 тыс. сайтов, что составляло примерно 0,1 % от общего числа сайтов в мире (либо около 0,3 % от числа активных сайтов в мире).

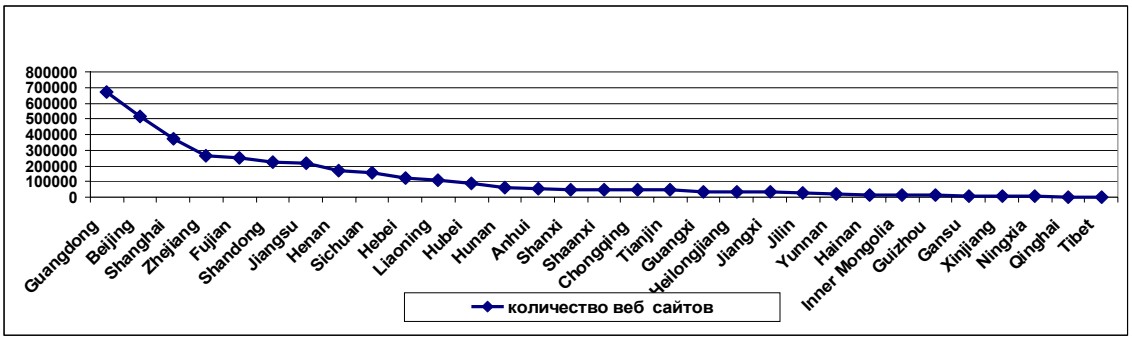

Рис. 4. Распределение количества веб-сайтов по регионам Китая

Количество веб-страниц. В [16, 17] приводятся данные, что в мае 2006 г. китайская поисковая система Baidu предоставляла своим пользователям доступ на основе индекса более 740 млн веб-страниц. По данным [5], общее количество веб-страниц в сегменте Интернета Китая в конце 2015 г. составляло 212,3 млрд. На рис. 5 представлено распределение количества веб-страниц по регионам Китая (подготовлено на основе данных работы [5], где из общего количества веб-страниц удалены дублирующие страницы). По оси абсцисс показаны регионы в порядке

убывания общего количества веб-страниц (верхний ряд). По оси ординат — количество веб-страниц. Из графика видно, что общее количество веб-страниц меняется от более чем 85 млрд для города Beijing (Пекин, население 15 810 000, до примерно 34 млн для провинции Qinghai (Цинхай, население 5 180 000). Под графиком общего количества веб-страниц — график соответствующего количества статических веб-страниц от более чем 50 млрд для Beijing до примерно 20 млн для Qinghai. Нижний график — соответствующее количество динамических веб-страниц (генерируемых программно в процессе выполнения запросов пользователей) от более чем 34 млрд для Beijing до примерно 13 млн для Qinghai. Общее количество веб-страниц по всей стране составляет более 212 млрд, из них более 131 млрд — это статические веб-страницы и более 80 млрд (примерно 38 %) — это динамические веб-страницы, при этом соотношение статических к динамическим веб-страницам по стране составляет 1,63. По отдельным регионам Китая соотношение статических к динамическим страницам меняется от 4,3 для города Chongqing (Чунцин), 3,19 для провинции Jiangsu (Цзянсу) до 0,37 и 0,5 для автономных районов Ningxia (Нинся-Хуэйский), Xinjiang (Синьцзян-Уйгурский).

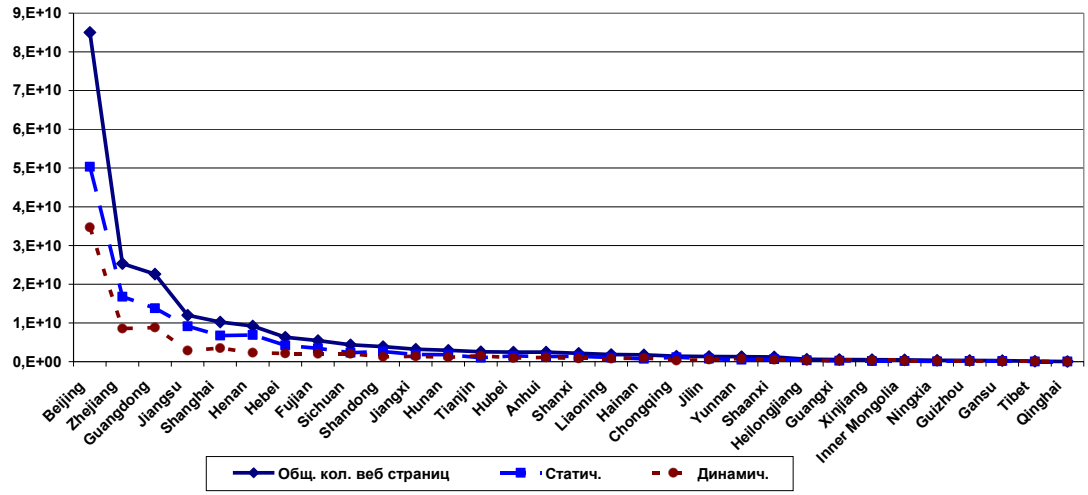

Рис. 5. Распределение количества веб-страниц по регионам Китая (приведено распределение общего количества страниц, а также распределение статических и динамических страниц)

Сравнить количество веб-страниц китайского контента с оценками объема мирового контента можно на основе данных [18–21]. В работах [18, 19] приводятся данные, в соответствии с которыми в сентябре 2016 г. количество проиндексированных поисковыми системами веб-страниц в мире составляет не менее 4,72 млрд страниц. В работе [20] количество проиндексированных веб-страниц в мире в 2005 г. оценивалось по крайней мере в 11,5 млрд страниц, а в работе [21] эта оценка в 2015 г. составляла 304,5 млрд. (Значительная разница в оценках количества проиндексированных веб-страниц в мире связана с отличиями методик оценки).

Периодичность обновления веб-страниц. На рис. 6 представлена периодичность обновления веб-страниц по регионам Китая (подготовлено с использованием данных работы [5]). По оси абсцисс показаны регионы в алфавитном порядке, а по оси ординат — проценты веб-страниц. Для каждого региона на графи-

ке представлены проценты веб-страниц, обновляемые еженедельно, ежемесячно, каждые 3 месяца, каждые 6 месяцев и реже, чем каждые 6 месяцев. Из графика видно, что максимальный процент веб-страниц, обновляемых еженедельно в провинции Gansu — 10,2 %, а реже, чем каждые 6 месяцев в провинции Hainan — 22,6 %. Средние значения для всех пяти показателей по Китаю составляют: 4,5 %, 24,4 %, 33,0 %, 27,6 % и 10,5 %.

Для сравнения обновляемости веб-страниц китайского контента с обновлением веб-страниц в мире соответствующие оценки были получены на основе интерфейса расширенного поиска системы Google https://www.google.ru/advanced_search (в поле поиска любых слов задавались запросы «the», «end» и др., а в поле сроков обновлений — от 24 часов до произвольного срока). Полученные результаты приведены на рис. 7.

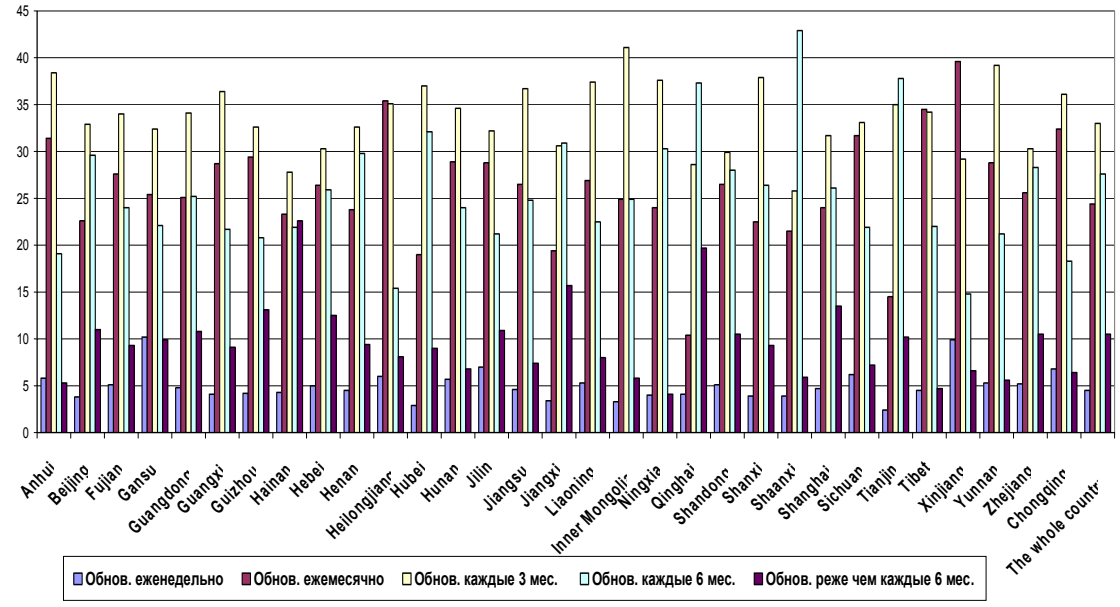

Рис. 6. Периодичность обновления веб-страниц по регионам Китая

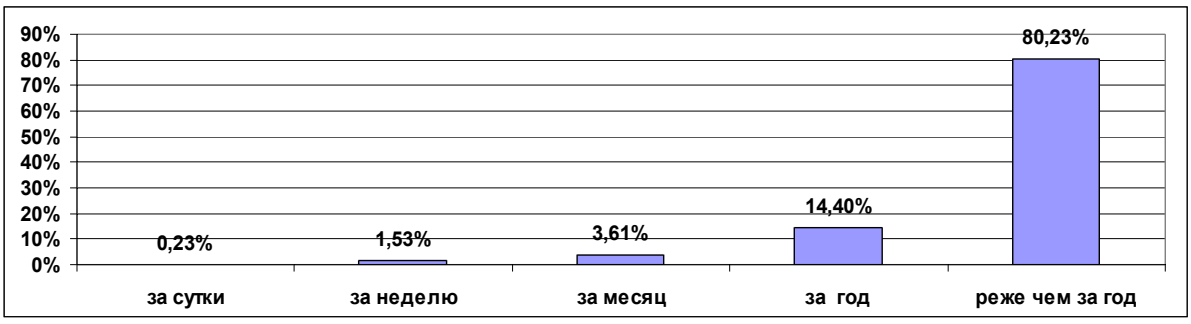

Рис. 7. Обновление веб-страниц сети Интернет на основе данных Google

По данным запросов к Google, из общего количества веб-страниц (более 25-ти млрд) за 24 часа обновляется около 0,23 % страниц, за неделю более 1,5 % и т.д., но более 80 % контента обновляется реже, чем за год.

Языки веб-страниц. На рис. 8 представлено использование языков и соответствующих кодировок при подготовке веб-страниц по каждому региону Китая (подготовлено с использованием данных из [5]). По оси абсцисс показаны регионы в алфавитном порядке, а по оси ординат — проценты веб-страниц. Для каждого региона на графике нижней частью колонки представлен процент веб-страниц, подготовленных на китайском языке, следующая над ней часть колонки отражает процент страниц, подготовленных на традиционном китайском языке, далее над ней часть колонки — процент страниц на английском языке и самая верхняя часть колонки — процент веб-страниц, подготовленных на других языках.

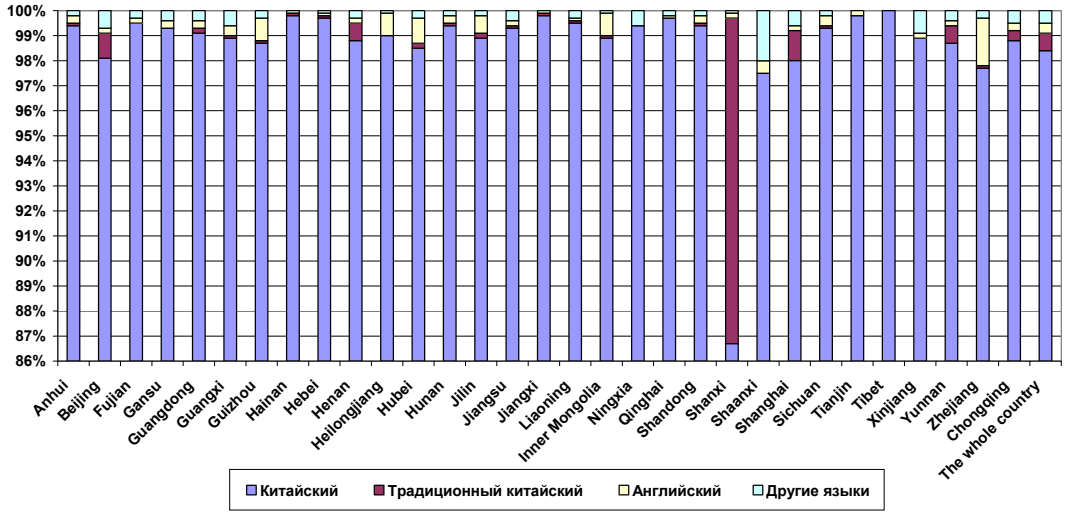

Рис. 8. Использование языков и соответствующих кодировок при подготовке веб-страниц по каждому региону Китая

Для сравнения представленных результатов по языкам на веб-ресурсах Китая с помощью средств расширенного поиска систем Google и Bing были получены соответствующие оценки для веб-ресурсов в доменах .cn и .com. Оценка количества веб-документов с помощью Google для упрощенного китайского языка показывает почти 0,5 млрд в домене .cn и около 100 млн в домене .com; для английского языка — около 20-ти млн в домене .cn и более 25-ти млрд в домене .com; для немецкого языка более 5-ти млн в домене .cn и более 3-х млрд в домене .com; для французского языка более 10-ти млн в домене .cn и почти 0,5 млрд в домене .com. Оценка количества веб-документов с помощью Bing для английского языка показывает более 7-ми млн в домене .cn и более 5-ти млрд в домене .com; для немецкого языка более 100 тыс. в домене .cn и около 200 млн в домене .com.

Форматы веб-страниц. Рис. 9 и 10 отражают использование форматов данных при создании веб-страниц и мультимедиа на сайтах Китая (подготовлено с использованием данных из [5]).

Сравнить форматы веб-страниц китайского контента с оценками объема мирового контента можно на основе данных [11, 22, 23]. По данным [22] в 2013 г. в мировом Интернете количество файлов формата pdf превышало более чем в 6 раз количество файлов формата doc и docx, а в сегменте .ru наоборот, количество файлов pdf было меньше количества файлов doc и docx. На рис. 11 в процентах при-

ведены оценки использования форматов pdf, doc/docx, rtf, txt при подготовке веб-документов. Показаны оценки использования форматов в мировом Интернете в 2013–2014 годах по данным [11] (обозначены мир 2013 и мир 2014), в доменах Германии и Японии в 2014 г. по данным [23] (обозначены .de 2014, .jp 2014), а также в доменах Германии, Японии и Китая в 2016 г. (обозначены .de 2016, .jp 2016 и .cn 2016), полученные с помощью средств расширенного поиска системы Google.

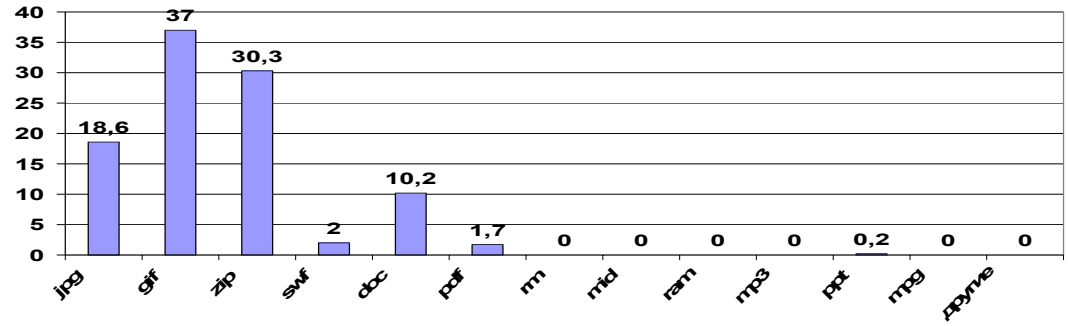

Рис. 9. Использование форматов данных при подготовке ресурсов мультимедиа на веб-страницах сайтов Китая (в процентах)

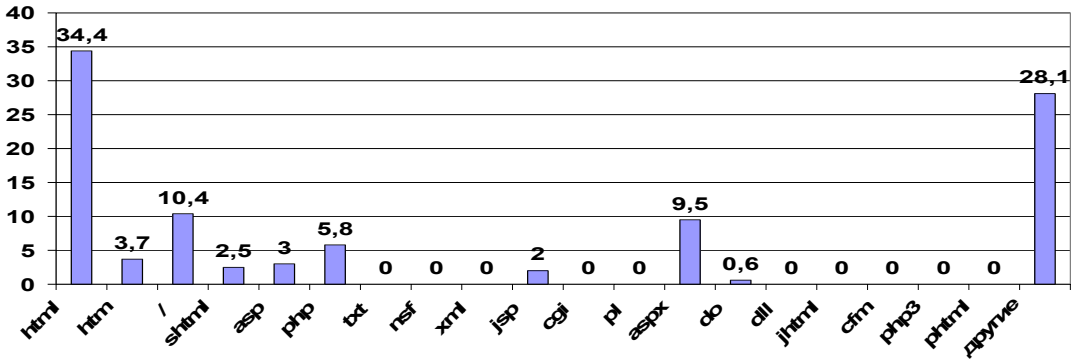

Рис. 10. Используемые форматы при подготовке веб-страниц на сайтах Китая (в процентах)

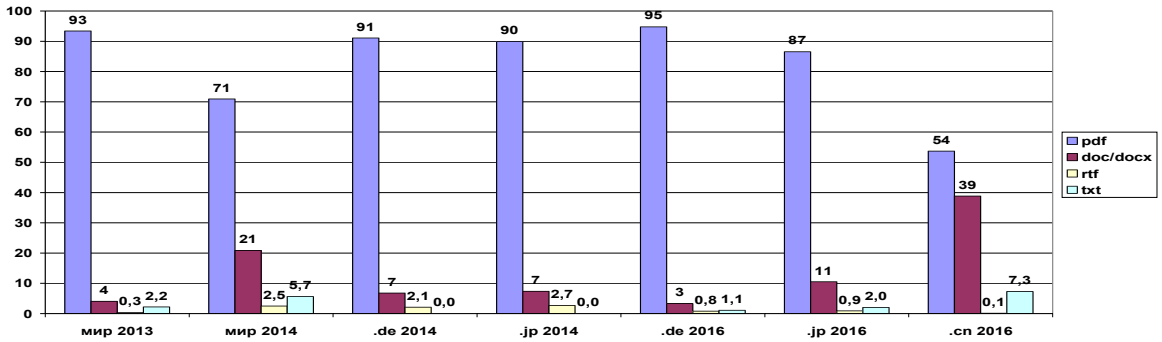

Рис. 11. Оценки использования форматов при подготовке веб-страниц в мировом Интернете (мир 2013 и мир 2014 [11]), в доменах .de 2014, .jp 2014 [23], а также в доменах .de 2016, .jp 2016 и .cn 2016, полученные с помощью средств расширенного поиска системы Google. Оценки приведены в процентах

Использование социальных сетей. Рис. 12 отражает процент пользователей Интернетом, использующих ресурсы различных китайских социальных сетей (подготовлено на основе данных работы [5]). По оси абсцисс представлены проценты, а по оси ординат — наиболее популярные китайские социальные сети.

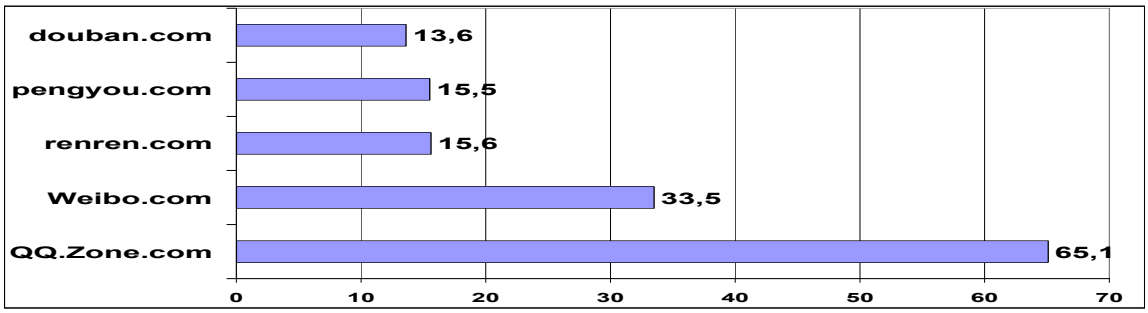

Рис. 12. Процент пользователей Интернетом, использующих ресурсы основных социальных сетей Китая

Оценки ряда характеристик этих сетей, полученные с помощью средств расширенного поиска систем Google и Bing, представлены на рис. 13. Для социальных сетей Weibo, Qzone, Renren, Pengyou, Douban оценки характеристик получены с помощью Google, а для Weibo и Douban — дополнительно с помощью Bing. Оценки приведены для таких характеристик (обозначения перечислены под графиком): количество веб-страниц, определяемое поисковой системой для социальной сети, количество веб-страниц на китайском упрощенном языке, китайском традиционном, английском, немецком, французском языке, а также количество веб-страниц в социальной сети за последние 24 часа, неделю, месяц, год. Соответствующие столбцы для каждой социальной сети приведены в указанном порядке. (Если характеристика не определена — то столбец отсутствует).

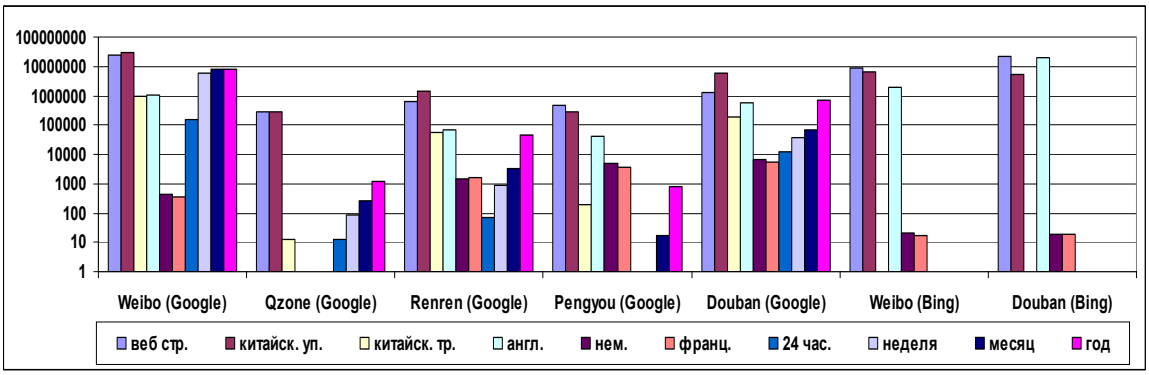

Рис. 13. Оценки характеристик китайских социальных сетей, полученные с помощью средств расширенного поиска систем Google и Bing

Ось ординат на рисунке представлена в логарифмическом формате, отражает оценку количества веб-страниц, определенную для соответствующей характеристики социальной сети. По оси абсцисс перечислены рассматриваемые социаль-

ные сети. Из рис. 13 видно, что общее число веб-страниц в этих сетях близко к количеству страниц на китайском упрощенном языке и насчитывает млн страниц.

Для сравнения, аналогичным способом были получены соответствующие характеристики социальных сетей Twitter и Facebook, которые представлены на рис. 14.

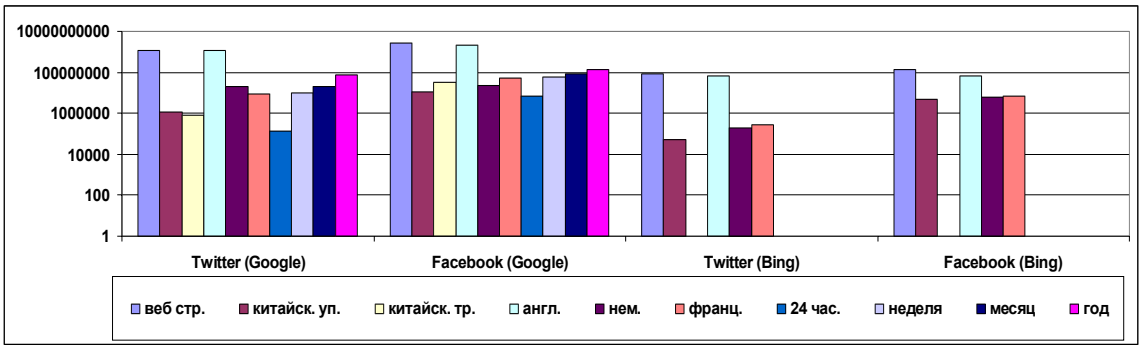

Рис. 14. Оценки характеристик социальных сетей Twitter и Facebook, полученные с помощью средств расширенного поиска систем Google и Bing

Обозначения на рис. 14 аналогичны обозначениям на рис. 13. Из рис. 14 видно, что количества веб-страниц в китайских социальных сетях оцениваются поисковыми системами в десятки млн, а в международных социальных сетях — в сотни млн.

Поисковая система Baidu. Baidu.com была основана в 2000 г. и в 2004 г. стала лидирующей поисковой системой в Китае. По количеству обрабатываемых запросов занимает 2 место в мире (с долей в глобальном поиске 18 %). В 2006 г. китайская поисковая система Baidu предоставляла своим пользователям доступ на основе индекса более 740 млн веб-страниц, 80 млн изображений и 10 млн файлов мультимедиа [16].

По данным за декабрь 2015 г. [5] в Китае насчитывается 566 млн пользователей поисковых систем (годовой рост 44 млн или 8,4 %, из них 478 млн мобильных поисковых пользователей, при общем количестве пользователей Интернетом в стране 688 млн). Поисковые системы являются вторым из наиболее используемых типов базовых приложений (после пересылки сообщений). По данным [24], во вторую половину 2014 г. Baidu использовали около 91,2 % пользователей поисковых систем (мобильные пользователи составили 90,3 %), далее следует поисковая система Soso/Sogou — использовали 45,8 % пользователей, поисковая система 360 — 38,6 % пользователей и Google — 27,4 % пользователей. На остальные китайские поисковые системы (Shenma, Easou, Youdao и др.) приходится от нескольких процентов до долей процентов пользователей.

Поисковая система Baidu, также как и системы Google, Bing, предоставляет пользователям средства расширенного поиска в виде поисковых операторов site, filetype, inurl и т.д., а также в виде специального интерфейса http://www.baidu.com/gaoji/advanced.html, где можно уточнить диапазон поиска, задав адрес конкретного сайта, формат файлов, язык, период времени и т.п.

Среди основных сервисов, которые предоставляет поисковая система Baidu, можно перечислить следующие (для их использования необходима регистрация в Baidu). Baidu Feng Yun Bang — самые обсуждаемые в Китае темы (здесь можно отбирать информацию по нужной тематике, определять, что популярно и в каком именно регионе). Baidu Zhidao — это сервис вопросов и ответов, подобный сервису Ответы.Mail.ru. Baidu Index — сервис, позволяющий анализировать поисковые запросы с выбранными ключевыми словами (насколько запрос был популярен в выбранном периоде: неделя, месяц, полгода и т.д.). Baidu Baike — это китайский аналог Википедии, в результатах поиска часто есть ссылка на статью в Baike. Baidu Tieba — это множество тематических сообществ для дискуссий, похожих на социальные сети. Baidu Webmaster — сервис, аналогичный сервисам Гугл.Вебмастер и Яндекс.Вебмастер.

Подробнее сервис Baidu Index рассмотрен в работе [25], посвященной анализу данных из запросов при поиске веб-ресурсов, где рассматриваются его возможности и отличия от аналогичного сервиса Google Trends. Пользователи вводят миллионы запросов в поисковые системы каждый день. Сервисы Baidu Index (index.baidu.com) и Google Trends (google.com/trends) создают отчеты о поисковых объемах (search volume) терминов, основываясь на запросах, которые пользователи ввели в поисковые системы.

Сервисы Baidu Index и Google Trends используются для кратковременного прогнозирования экономической деятельности, определения эпидемий болезней, также показана корреляция между качеством кинофильмов и поисковыми объемами данных кинофильмов и т. д. При использовании сервиса Baidu Index формируются отчеты с абсолютными значениями поисковых объемов, а Google Trends — с относительными значениями. В [25] было отобрано 50 китайских университетов из годового рейтинга топ-50 и 50 китайских ИТ-компаний также из соответствующего рейтинга. Для каждого названия университета и компании из сервисов Baidu Index и Google Trends были найдены значения поисковых объемов. С их помощью была обнаружена корреляция между рейтингами организаций и их поисковыми объемами, а также корреляция между значениями поисковых объемов из Baidu Index и Google Trends.

Учитывая показанную в [25] корреляцию между рейтингами организаций и их поисковыми объемами, для данного обзора контента при помощи рассмотренных сервисов были получены графики изменения поисковых объемов для китайских социальных сетей Qzone, Weibo, Renren, Pengyou и Douban на территории Китая за последние 5 лет. Наибольшее значение поискового объема показывает график Weibo, наименьшее — график Pengyou. Графики поисковых объемов для Qzone, Renren и Douban занимают промежуточное положение между этими двумя графиками. В относительных единицах полученное среднее значение показателя поискового объема составило: для Weibo — 52, Douban — 17, Renren — 15, Qzone — 11 и Pengyou — 1. По данным сервиса территориально южным районам Китая соответствуют большие поисковые объемы для Weibo, северо-западным и частично центральным районам — большие поисковые объемы для Qzone, северо-восточным и центральным районам соответствуют большие поисковые объемы для Renren. Для сравнения, вместо Pengyou и Douban в рассматриваемых сервисах также задавались американские социальные сети Facebook и Twitter. Было получено сле-

дующее соотношение средних значений показателей поискового объема: для Facebook — 32, Twitter — 12, Weibo — 5, Douban — 2, Renren, Qzone — около 1.

В 2003 году в Китае была введена в эксплуатацию система фильтрации содержимого Интернет — проект «Золотой щит». Это китайский фаервол, система серверов на интернет-канале между провайдерами и международными сетями передачи данных, которая фильтрует информацию. В результате, ограничение доступа коснулось социальных сетей Twitter, Facebook, Google+ и других ресурсов [26].

Возможности сбора контента

RSS (Rich Site Summary — обогащенная сводка сайта) — это семейство XML-форматов, используемое для публикации и доставки часто изменяющейся информации (заголовков новостей, анонсов статей, новых записей в блогах и т.д.), это технология, которую применяют пользователи Интернетом для получения обновлений с интересующих их веб-страниц. Природа RSS обусловила то, что одним из эффективных способов сбора контента информационных ресурсов Интернета является использование каналов RSS. В 2005 году около 30 % владельцев медиасайтов предоставляли контент через RSS-каналы [27]. К 2008 году это число возросло до 50 %.

Количество каналов RSS в 2004 г. составляло около 307 тыс., в 2016 г. директория Feedage.com (в которой RSS-каналы со всего мира представлены 15 категориями с возможностью поиска) объединяет более 3,1 млрд каналов.

Наличие RSS-каналов обеспечивает непрерывное получение обновлений веб-сайта как заинтересованными пользователями, так и автоматическими системами анализа веб-ресурсов. В частности, авторами разработана система мониторинга веб-ресурсов, базирующаяся на использовании RSS-каналов (рис. 15).

Рис. 15. Интерфейс эксперта-аналитика подключения RSS-канала к системе мониторинга веб-сайтов

Анализ использования RSS-каналов на веб-ресурсах китайского сегмента Интернета и в мире показывает следующее. В работе [28] исследовано использование технологий Web 2.0 (социальных сетей, технологий wiki, блогов, RSS-каналов, обмен мгновенными сообщениями и функции каталогизации) в библиотеках 38-ми ведущих университетов Китая. Показано, что RSS является второй по частоте использования технологией (представлена в 55 % университетских библиотек). Отмечаются три основные цели использования RSS в библиотеках китайских университетов: уведомление об информации, представляющей интерес для читателей по инициативе библиотеки — новости и события библиотеки, доступность новых книг, т.е. информационная база данных; уведомление о личной информации про пользование библиотекой; синдикация тематической информации для легкого и своевременного доступа. Эти цели предполагают разные уровни технологической поддержки, поэтому большинство библиотек обеспечивают в основном базовые возможности RSS-каналов. Только RSS-каналы библиотек Шанхайского университета ориентированы на достижение всех трех целей. В работе [29] рассмотрено внедрение технологий Web 2.0, в том числе и RSS-каналов, в библиотеках 30-ти ведущих университетов Китая. Показано, что из всех технологий Web 2.0, каналы RSS получили наибольшее распространение в библиотечных проектах (далее следует передача сообщений, использование блогов и т.д.). Больше всего каналы RSS используются для распространения новостей и уведомлений — в 12-ти университетах из 30-ти, что составляет 43 %.

В работе [30] исследуется использование приложений Web 2.0, в том числе и RSS-каналов (распространение информации библиотек для пользователей), на сайтах 120-ти крупнейших библиотек трех регионов — Северной Америки, Европы и Азии. В числе 40-ка крупнейших библиотек региона Азии рассматривались Гонконгская публичная библиотека и библиотеки Китайского университета Гонконга, а также Гонконгского университета науки и технологий, также библиотека Университета Цинхуа, Национальная центральная библиотека (Тайвань) и Национальная библиотека Китая. Из общего количества проанализированных 120-ти сайтов библиотек, распространение информации с помощью каналов RSS применяется на 28-ми сайтах университетов Северной Америки (что составляет около 70 %), 17-ти и 15-ти сайтах университетов Европы и Азии (43 % и 37 % соответственно). В целом, исследование показало, что RSS-каналы используются примерно в 50 % крупнейших библиотек трех регионов и занимают второе место по популярности среди приложений Web 2.0 после блогов (примерно 57 %). Также для сравнения, данные по использованию каналов RSS в 100-та ведущих академических библиотеках США приводятся в [31]. По данным этого исследования, из Web 2.0 технологий больше всего используются социальные сети — в 100 % библиотек. Блоги используются в 99 %, а RSS-каналы в 97 % исследованных академических библиотек США.

Результаты анализа использования RSS-каналов на веб-ресурсах китайского сегмента Интернета и в мире приведены на рис. 16. Рисунок показывает, что около половины сайтов библиотек используют каналы RSS, это больше чем в среднем по странам Азии, но меньше чем в странах Европы и США.

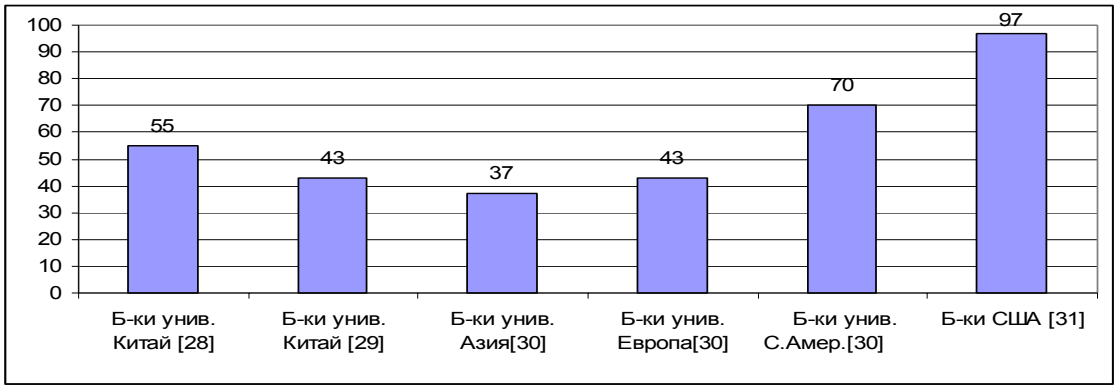

Рис. 16. Анализ использования RSS-каналов в составе Web 2.0 технологий на сайтах библиотек ведущих университетов Китая [28, 29], библиотек ведущих университетов стран Азии, Европы, Сев. Америки [30], а также академических библиотек США [31]. Показан процент сайтов библиотек с использованием RSS-каналов от общего числа исследованных в указанной работе библиотек

Основные категории контента. Для оценки возможностей сбора контента китайского сегмента Интернета с использованием каналов RSS были выделены следующие категории веб-ресурсов: порталы газет; новостные порталы; сайты университетов и институциональные репозитории; сайты госучреждений; сайты правовой информации. По каждой категории веб-ресурсов было отобрано около двадцати сайтов с учетом рейтинга Alexa.com (который вычисляется на основе оценки посещаемости ресурсов Интернета), а также каталога 4 International Media & Newspapers (4imn.com, его рейтинг учитывает кроме Alexa.com еще Google Page Rank и Majestic Seo показатели). Сайты многих ведущих газет и новостных порталов, также как и рассмотренные в работах [28–31] сайты крупнейших библиотек, для распространения информации используют RSS-каналы. Анализ сайтов китайских газет показывает, что около 40 % китайскоязычных версий сайтов и более 50 % англоязычных версий сайтов используют RSS-каналы для распространения информации. Около 60 % китайскоязычных версий новостных порталов и более 70 % англоязычных версий порталов также используют RSS-каналы для распространения информации.

Кроме перечисленных категорий веб-ресурсов, для оценки возможностей сбора контента китайского сегмента Интернета рассматривались характеристики основных китайских социальных сетей (см. таблицу). К числу наиболее популярных социальных сетей и блогов относят следующие [1, 5]. Sina Weibo (микроблог) является эквивалентом социальной сети Twitter в Китае (weibo.com). Douban (douban.com) представляет отзывы пользователей и рекомендации для книг, музыки и фильмов. В Qzone.qq (qzone.com) около 150-ти млн пользователей обновляют свои страницы не реже одного раза в месяц, что делает QZone одним из наиболее активных сайтов. Renren (renren.com) был создан для студентов высших учебных заведений. Сайт идентичен Facebook и имеет во многом тот же пользовательский интерфейс, инструменты и функции.

Сравнительные характеристики основных китайских социальных сетей и сетей Facebook, Twitter на основе показателей Alexa.com

	Facebook	Twitter	Weibo	Qzone.qq	Douban	RenRen
Рейтинг Alexa.com	3 gl; 2 US	8 gl; 8 US	20 gl; 5 Cn	10 gl; 2 Cn	868 gl; 72 Cn	1783 gl; 195 Cn
Год создания, краткие характеристики	2004 г. — более 1 млрд пользователей	2006 г., — более 200 млн пользователей	2009 г. — 250 млн аккаунтов, 90 млн постов/день	2005 г. — более 600 млн пользователей	2005 г. — около 200 млн. пользователей	2005 г. — более 160 млн пользователей
Страны пользователей	22 % — US 8 % — In 4 % — Br 3 % — GB 3 % — Gr	22 % — US 14 % — Jp 7 % — In 6 % — GB 4 % — Mx	97 % — Cn 0,7 % — US 0,6 % — Tw	98 % — Cn 0,5 % — US	92 % — Cn 3 % — US 0,8 % — Hk 0,8 % — Tw	92 % — Cn 4 % — US 0,8 % — Tw 0,6 % — Jp 0,5 % — Hk

Анализ контента с использованием возможностей weibo.com. В работе [2] проведен сравнительный анализ трендов в китайских социальных сетях на основе списка 50-ти ключевых слов, которые появляются чаще всего в твитах пользователей weibo.com (ранжируются по частоте появлений за последний час) и анализа трендов тем в Твиттере. Показано, что среднее время нахождения каждого ключевого слова в списке трендинга составляет около 6-ти часов. Кроме того, распределение количества часов нахождения каждой темы в списке трендинга соответствует степенному закону. Степенной характер распределения показывает, что только нескольким темам свойственна долговременная популярность. Другой результат состоит в том, что большинство ключевых слов исчезает из списка трендинга после определенного количества раз появлений и пропаданий. Определено, что распределение количества повторных появлений ключевых слов в списке трендинга также близко к степенному закону. Полученные выводы подобны результатам анализа трендов тем в Твиттере [2]. Важное отличие состоит в том, что среднее время трендинга в weibo.com значительно выше (около 6-ти часов по сравнению с 20–40 минутами в Твиттере). Это говорит о том, что weibo.com не может иметь так много конкурирующих тем, как Твиттер.

Для получения дополнительных характеристик веб-контента на основе трендов weibo.com, при подготовке данного обзора также выполнялся мониторинг ключевых слов из топ-50 списка вейбо. С этой целью было выбрано приложение для браузера Firefox — Alertbox, которое было настроено для мониторинга списка ключевых слов с периодом около 1 часа.

В результате почти пятидневного мониторинга были получены примеры графиков изменения количества страниц weibo.com, содержащих определенные ключевые слова из списка топ-50. Например, ключевое слово pokemongo на страницах Вейбо с 14:00 22.07.2016 до 4:00 23.07.2016 (в течение около 14-ти часов) занимало с 18-го по 9-е места рейтинга топ-50 с количествами страниц примерно от 10-ти тыс. до 160-ти тыс. (на рис. 17 нижний график, обозначение ключевое

слово 1). Ключевое слово 快乐大本营 (Счастливый лагерь — популярное развлекательное шоу Китая) на страницах Вейбо с 15:00 23.07.2016 до 9:00 24.07.2016 (около 18-ти часов) занимало с 47-го по 2-е места рейтинга топ-50 с количествами страниц примерно от 20-ти тыс. до более чем 500 тыс. (на рис. 17 верхний график, обозначение ключевое слово 2). На рис. 17 по оси абсцисс показаны номера сканирования страницы топ-50 Вейбо, а по оси ординат — количество страниц социальной сети Вейбо с соответствующими ключевыми словами.

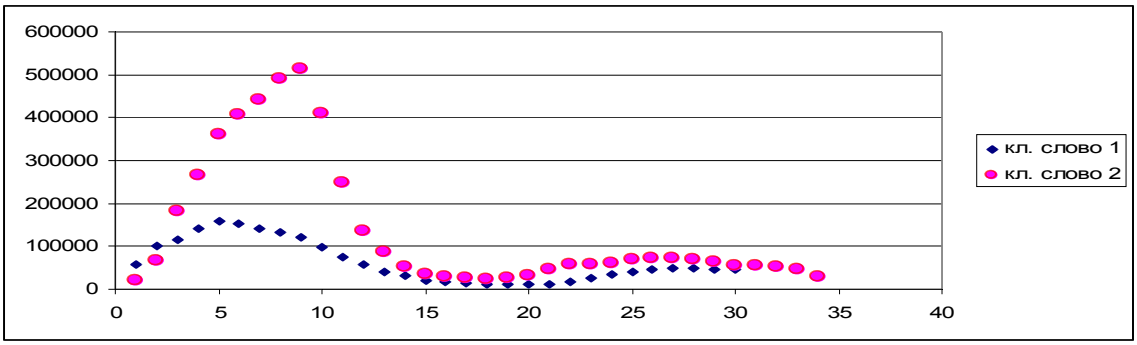

Рис. 17. Анализ количества появлений ключевых слов pokemongo (ключевое слово 1) и 快乐大本营 (ключевое слово 2 — Счастливый лагерь) в твитах социальной сети Вейбо на основе мониторинга страницы топ-50 Вейбо

Для сравнения, анализировались изменения поисковых объемов для ключевых слов pokemongo и 快乐大本营 (Счастливый лагерь), получаемые на основе сервисов поисковых систем Baidu Index и Google Trends, рассмотренные выше. На рис. 18 приведены изменения поисковых объемов для этих ключевых слов за последние 12 месяцев (по оси абсцисс) в относительных единицах (по оси ординат) для Китая (использован скриншот интерфейса www.google.com.hk/trends/). Нижний график на рисунке имеет пик до 100 единиц (соответствует дате 10.07.2016) и показывает изменение поискового объема для ключевого слова pokemongo. Верхний график на рисунке значительных пиков не имеет и показывает изменение поискового объема для ключевого слова 快乐大本营 (Счастливый лагерь).

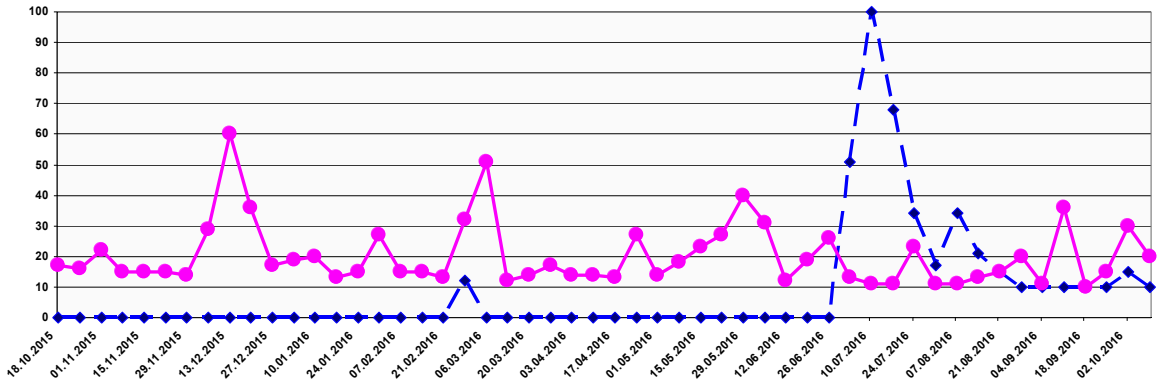

Рис. 18. Анализ изменения поискового объема для ключевых слов pokemongo и 快乐大本营 (Счастливый лагерь) в течение последнего года

Сравнение данных по графикам на рис. 17, 18 показывает близость между датами пика поискового объема для pokemongo (10.07.2016) и нахождением этого ключевого слова в топ-50 страниц Вейбо (22–23.07.2016). На основе данных мониторинга ключевых слов также была получена оценка количества ключевых слов из топ-50, появляющихся в твитах Вейбо за сутки, которая составила 100–150 слов.

Выводы

Рассмотрены особенности контента китайского сегмента Интернета, которые необходимо учитывать при его сборе: количества веб-страниц и веб-сайтов, периодичность обновления и языки страниц по регионам Китая; форматы веб-страниц, оценки популярности порталов газет, новостей, данные об использовании китайских соц. сетей и другие характеристики. К особенностям китайского сегмента веб-пространства следует отнести:

— темпы роста количества веб-ресурсов и числа пользователей превосходят всемирный сегмент Интернета;

— наличие собственных социальных сетей, объемы которых превосходят объемы аналогичных в мире;

— наличие собственной основной поисковой системы Baidu (наряду с еще несколькими), ориентированной преимущественно на китайский язык (существенной проблемой является применение латиницы и кириллических кодов) и покрывающей значительную часть веб-ресурсов китайского сегмента Интернета;

— пока еще относительно небольшое представление ресурсов в формате RSS (связанное с некоторым запаздыванием внедрения интернет-технологий). Вместе с тем представление веб-ресурсов в RSS-формате в настоящее время возрастает и все шире используется в мобильных приложениях;

Показаны возможности сбора контента с использованием каналов RSS и программных средств системы мониторинга веб-ресурсов, приведены данные о внедрении каналов RSS в составе Web 2.0-технологий на сайтах китайских университетов, газет, новостных порталах и т.д., показана возможность мониторинга ключевых слов на weibo.com для анализа новостного контента.

1. *Deans P.C.* A framework to understanding social media trends in China / P.C. Deans, J.B. Miles // The 11-th Internation. DSI and APDSI Joint Meeting, Taipei, Taiwan. — 2011, July 12-16. — P. 12–16.

2. *Yu L.* Dynamics of trends and attention in chinese social media / L. Yu, S. Asur, B.A. Huberman // arXiv preprint arXiv:1312.0649, 2013. — P. 1–17.

3. *Bolsover G.* Social Foundations of the Internet in China and the New Internet World: A Cross-National Comparative Perspective / G. Bolsover, W.H. Dutton, G. Law. — Oxford Internet Institute, University of Oxford,. — 2013. — P. 1–22.

4. *Internet* Users by Country (2016) [Электронный ресурс]. — Режим доступа: http://www.internetlivestats.com/internet-users-by-country/. — Название с экрана.

5. *CNNIC*. (2016) // The 37-th Statistical Report on Internet Development in China.

6. *Vaughan L.* Equal representation by search engines? A comparison of websites across countries and domains / Liwen Vaughan, Zhang Yanjun // Journal of Computer-Mediated Communication 12.3. — 2007. — P. 888–909.

7. *Vaughan L.* Search engine coverage bias: evidence and possible causes / Liwen Vaughan, Mike Thelwall // Information processing & management 40.4. — 2004. — P. 693–707.

8. *Orduña-Malea E.* The dark side of Open Access in Google and Google Scholar: the case of Latin-American repositories / E. Orduña-Malea, E. Delgado-López-Cózar // Scientometrics 102.1. — 2015. — P. 829–846.

9. *Orduña-Malea E.* Methods for estimating the size of Google Scholar / E. Orduña-Malea // Scientometrics 104.3. — 2015. — P. 931–949.

10. *Arlitsch K.* Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar / K Arlitsch, P.S. O'Brien // Library Hi Tech 30.1. — 2012. — P. 60–81.

11. *Ланде Д.В.* Підхід до оцінки живучості наукових публікацій при довготерміновому зберіганні в інтернет-середовищі / Д.В. Ланде, Б.О. Березін // Реєстрація, зберігання і оброб. даних, — 2014. — Т. 16, № 4. — С. 34–43.

12. *Number* of Chinese websites nearly halves [Электронный ресурс]. — Режим доступа: http://www.telegraph.co.uk/news/worldnews/asia/china/8634534/Number-of-Chinese-websites-nearly-halves.html. — Название с экрана.

13. *Content* Delivery Network [Электронный ресурс]. — Режим доступа: http://www.internet-livestats.com/. — Название с экрана.

14. *August* 2016 Web Server Survey [Электронный ресурс]. — Режим доступа: https://news.netcraft.com/archives/category/web-server-survey/. — Название с экрана.

15. *Сколько* сайтов в Уанете? [Электронный ресурс]. — Режим доступа: http://vlasti.net/news/166418. — Название с экрана.

16. *Baidu.com*, Inc. (ADR): Company Report [Электронный ресурс]. — Режим доступа: https://web.archive.org/web/20060501050256/http://moneycentral.msn.com/investor/research/profile.asp?Symbol=BIDU. — Название с экрана.

17. *Baidu* [Электронный ресурс]. — Режим доступа: https://en.wikipedia.org/wiki/Baidu. — Название с экрана.

18. *The size* of the World Wide Web (The Internet) [Электронный ресурс]. — Режим доступа: http://www.worldwidewebsize.com. — Название с экрана.

19. *Bosch A.* Estimating search engine index size variability: a 9-year longitudinal study / A. Bosch, T. Bogers, M. Kunder // Scientometrics. — 2016. — **107**(2). — P. 839–856.

20. *Gulli A.* The indexable web is more than 11.5 billion pages / Gulli, Antonio, and Alessio Signorini // Special interest tracks and posters of the 14th international conference on World Wide Web. ACM, 2005. — P. 902–903.

21. *Svabensky V.* Web Science 2015 Project 1: Web Size [Электронный ресурс] / V. Svabensky. — Режим доступа: http://www.fi.muni.cz/~xsvabens/erasmus/WSProject1.pdf. — Название с экрана.

22. *Статистика* распространенности файловых форматов в Интернете [Электронный ресурс]. — Режим доступа: http://rusrim.blogspot.ru/2013/07/blog-post_16.html. — Название с экрана.

23. *98 % of* .com is HTML but 38 % of .gov is PDF! [Электронный ресурс]. — Режим доступа: http://duff-johnson.com/2014/03/10/98-percent-of-dot-com-is-html-but-38-percent-of-dot-gov-is-pdf/. — Название с экрана.

24. *The 35-th* Statistical Report on Internet Development in China, 2014.

25. *Vaughan L.* Data mining from web search queries: A comparison of google trends and baidu index / Liwen Vaughan, Yue Chen // Journal of the Association for Information Science and Technology 66.1. — 2015. — P. 13–22.

26. Кого же блокирует китайский фаервол? [Электронный ресурс]. — Режим доступа: https://habrahabr.ru/post/238379/. — Название с экрана.

27. *Ma D.* Use of RSS feeds to push online content to users / D. Ma // Decision Support Systems. — 2012. — **54.1**. — P. 740–749.

28. *Han Z.* Web 2.0 applications in top Chinese university libraries / Zhiping Han, Yan Quan Liu // Library Hi Tech. — 2010. — **28.1**. — P. 41–62.

29. *Chua A.* A study of Web 2.0 applications in library websites / AYK Chua, DH Goh // Library & information science research. — 2010. — **32.3**. — P. 203–211.

30. *Si L.* An investigation and analysis of the application of Web 2.0 in Chinese university libraries / L. Si, R. Shi, B. Chen // The electronic library 29.5. — 2011. — P. 651–668.

31. *Boateng F.* Web 2.0 applications' usage and trends in top US academic libraries / F. Boateng, Y. Quan Liu // Library Hi Tech 32.1. — 2014. — P. 120–138.

Поступила в редакцию 19.09.2016