ПОСТРОЕНИЕ МОДЕЛИ ИНФОРМАЦИОННОГО СЕРВИСА НА БАЗЕ НАЦИОНАЛЬНОГО СЕГМЕНТА ИНТЕРНЕТ

Д. Ландэ[1], Б. Березин[1], О. Павленко[2],
[1]Институт проблем регистрации информации НАН Украины,
[2]Университет "Украина"

Актуальность и постановка проблемы

В настоящее время китайский сегмент Интернета является наибольшим в мире по количеству пользователей – более 688 млн. (что составляет более 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 [1–7] 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; собственные социальные сети и поисковые системы.

Контент китайского сегмента Интернета представлен 4,23 млн. веб-сайтов и 212,3 млрд 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 и других онлайн переводчиков позволяют преодолеть языковый барьер и повышают 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 исследованы и использованы недостаточно. Имеющиеся работы [1–5] и др., как правило, посвящены лишь отдельным характеристикам контента. В данной работе, на примере китайского сегмента сетевого информационного пространства показаны особенности контента национального сегмента Интернет. Предложено построение модели

информационного сервиса, обеспечивающего сбор контента китайского сегмента с помощью системы мониторинга веб-ресурсов на основе использования RSS-каналов, а также мониторинг ключевых слов социальной сети. Система мониторинга на основе каналов RSS является универсальной и может быть адаптирована к любому национальному сегменту Интернет.

Особенности контента китайского сегмента Интернет

Кроме высоких темпов роста количества веб-ресурсов и числа пользователей, китайский сегмент Интернета выделяется в мировой Сети наличием собственных социальных сетей, объемы которых соизмеримы с объемами аналогичных мировых; наличием собственной основной поисковой системы Baidu, ориентированной преимущественно на китайский язык и покрывающей значительную часть веб-ресурсов китайского сегмента Интернета; пока еще относительно небольшим, но развивающимся представлением ресурсов в формате RSS.

Социальные сети. Пользователи китайского сегмента Интернет являются активными участниками китайских социальных сетей. По данным [5], на конец 2015 г более 65 % пользователей Интернета использовали QQ.Zone.com, более 33 % – Weibo.com, около 15 % – использовали RenRen.com, Pengyou.com и Douban.com.

С помощью сервиса Alexa.com был проведен сравнительный анализ рейтинга и пользователей основных международных и китайских социальных сетей (таб. 1). Результаты подтверждают, что соизмеримые по размеру с Facebook и Twitter китайские сети по составу пользователей являются национальными сетями.

В работе [2] проведен сравнительный анализ трендов в китайских социальных сетях на основе списка 50 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 (это показывает, что только нескольким темам свойственна долговременная популярность).

Поисковая система Baidu. Baidu.com была основана в 2000 г. и в 2004г. 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.

Таблица 1.

	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

Среди значительного числа сервисов, предоставляемых Baidu, возможности созданного в 2014 г. сервиса http://xueshu.baidu.com, который называют Baidu Scholar, во многом аналогичны сервису Google Scholar, который существует с 2004 г. Сравнительный анализ особенностей представления научных публикаций украинских авторов в результатах, формируемых поисковыми системами Baidu Scholar и Google Scholar, отражен в таблице. Таблица содержит результаты поиска англо-, украино- и русскоязычных научных публикаций по инициалам и фамилии авторов, перечисленных в таблице. (Приводятся результаты поиска для варианта инициалов без разделения точкой и с разделением. Другие варианты написания не рассматривались). В столбцах Baidu Scholar и Google Scholar каждому автору, по которым выполнялся поиск, соответствует количество полученных в результате поиска ссылок. Для некоторых из полученных в результате поиска количеств ссылок, с целью детализации, через дробь приводится также второе число – количество реально отобранных из них ссылок, соответствующих запросу. В случаях больших количеств полученных при поиске ссылок (более 150), отбор ссылок, соответствующих запросам, производился из первых 150 полученных результатов. При представлении результатов поиска в виде дроби, по соотношению числителя и знаменателя может оцениваться качество поиска. Для детализации результатов, кроме обычного поиска по всем доменам, в

Google Scholar также производился поиск отдельно по доменам .org, .ua, и .cn. Из результатов поиска и таблицы видно, что англоязычные публикации сосредоточены на серверах домена .org (например, arxiv.org), а украино- и русскоязычные на серверах домена .ua (nbuv.gov.ua и др.).

Сравнение результатов использования Baidu и Google показывает близость полученных ответов на запросы для англоязычных публикаций в обоих системах. (Число полученных в результате поиска ссылок в столбце 3 таблицы незначительно превышает число полученных ссылок в столбце 4, при этом число предлагаемых ссылок не намного больше количества реальных публикаций, которые могут быть по ссылкам отобраны. Для большинства показанных примеров, из нескольких десятков полученных ссылок могут быть отобраны десятки реальных публикаций). Для украино- и русскоязычных публикаций наоборот, получена значительная разница в ответах Baidu и Google по сравнению с реальным числом публикаций. (Для Baidu, в отличие от Google, число предлагаемых ссылок значительно превышает количество реальных публикаций, которые могут быть по ссылкам отобраны. Из нескольких десятков тысяч ссылок для украино- и русскоязычных публикаций, реально могут быть отобраны из первых 150 ссылок несколько десятков публикаций, соответствующих запросам).

RSS-каналы. 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 Интернета для получения обновлений с интересующих их веб-страниц. Природа RSS обусловила то, что одним из эффективных способов сбора контента информационных ресурсов Итернет является использование каналов RSS. 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.

Анализ использования RSS-каналов на веб-ресурсах китайского сегмента Интернета и в мире показывает следующее. В работе [8] исследовано использование технологий Web 2.0 (социальных сетей, технологий wiki, блогов, RSS-каналов, обмен мгновенными сообщениями и функции каталогизации) в библиотеках 38 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 книги, 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 [9] 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 [10] 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-ти % 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 [11]. 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.

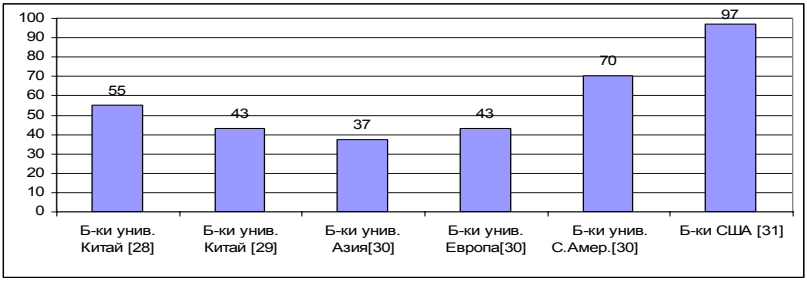

Рисунок 1 – Анализ использования RSS-каналов в составе Web 2.0 технологий на сайтах библиотек ведущих университетов Китая [8,9], библиотек ведущих университетов стран Азии, Европы, Сев. Америки [10], а также академических библиотек США [11]. Показан процент сайтов библиотек с использованием RSS-каналов от общего числа исследованных в указанной работе библиотек

Построение модели информационного сервиса

Предлагаемая модель информационного сервиса разрабатывается с учетом особенностей контента китайского сегмента Интернет [6,7]. Наличие RSS-каналов обеспечивает непрерывное получение обновлений веб-сайта как заинтересованными пользователями, так и автоматическими системами анализа веб-ресурсов. В частности, была разработана система мониторинга веб-ресурсов, базирующаяся на использовании RSS-каналов (рис. 2).

Система мониторинга веб-сайтов на основе использования каналов RSS обеспечивает сбор необходимого контента различной направленности, является универсальной и может быть с минимальными перестройками адаптирована к любому национальному сегменту Интернет.

Мониторинг ключевых слов. Кроме системы мониторинга веб-ресурсов на основе использования RSS-каналов, модель информационного сервиса предусматривает мониторинг социальной

сети Weibo. С целью мониторинга топ-50 ключевых слов сети Weibo было выбрано приложение для браузера Firefox – Alertbox, которое было настроено для мониторинга списка ключевых слов с периодом около 1 часа. В результате почти недельного мониторинга были получены примеры графиков изменения количества страниц weibo.com, содержащих определенные ключевые слова из списка топ-50.

Рисунок 2 – Интерфейс эксперта-аналитика подключения RSS-канала к системе мониторинга веб-сайтов

Например, ключевое слово pokemongo на страницах Вейбо 23.07.16 (в течение около 14 часов) занимало с 18 по 9 места рейтинга топ-50 с количествами страниц примерно от 10-ти тыс. до 160-ти тыс. (На рис. 3 нижний график, обозначение ключевое слово 1). Ключевое слово 快乐大本营 (Счастливый лагерь – популярное развлекательное шоу Китая) на страницах Вейбо с 23.07.16 до 24.07.16 (около 18-ти часов) занимало с 47 по 2 места рейтинга топ-50 с количествами страниц примерно от 20-ти тыс. до более 500 тыс. (На рис. 3 верхний график, обозначение ключевое слово 2).

На рис. 3 по оси абсцисс показаны номера сканирования страницы топ-50 Вейбо, а по оси ординат – количество страниц социальной сети Вейбо с соответствующими ключевыми словами.

Для более информативного мониторинга ключевых слов Weibo, целесообразно вместе с китайскими ключевыми словами отображать их перевод на английский, украинский, русский и др. языки. Примеры перевода ключевых слов с помощью Google, Yandex, Bing переводчиков приведены в табл. 3.

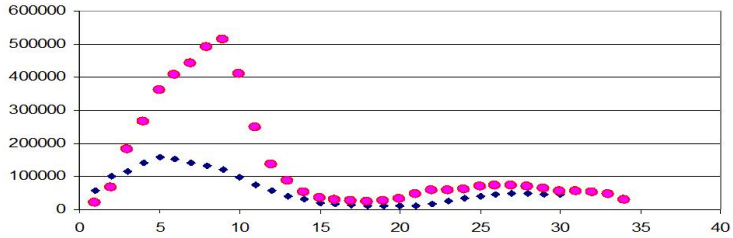

Рисунок 3. – Анализ количества появлений ключевых слов pokemongo (кл. слово 1) и 快乐大本营 (кл. слово 2 — Счастливый лагерь) в твитах социальной сети Вейбо на основе мониторинга страницы топ-50 Вейбо

С целью такого мониторинга используются три модуля на языке Perl: модуль выборки ключевых слов из веб-страницы, модуль обмена с API системы перевода (для получения соответствующих переводов ключевых слов с китайского на английский и русский языки), модуль представления ключевых слов вместе с переводами на веб-странице.

Таблица 3. Примеры онлайн перевода ключевых слов Weibo системами Google, Yandex, Bing

Ключ. слово	Рейтинг	Google	Yandex	Bing	Содержание
林依晨哭了	302947	Ariel крик	Ариэль плакала	Ариэль Крик	Актриса о награде
真的有蓝瘦的香菇	285822	Действительно тонкие голубые грибы	Правда синий тощие грибы	True blue тонкий гриб	Специалист о необычных грибах
你绑定的支付宝怎么办	285085	Вы связываете Alipay, как это сделать	Вы не привязаны к PayPal?	Какие привязки платежной карты вы	О переходе на новую систему оплаты
同班同学联名举报	124998	Совместный отчет Classmate	Одноклассница совместный доклад	Группы одноклассников сообщили	Одноклассники выступили против плагиата

Выводы

К особенностям китайского сегмента веб-пространства следует отнести:

– темпы роста веб-ресурсов и числа пользователей, которые по ряду характеристик превосходят всемирный сегмент Интернета;

– наличие собственных национальных социальных сетей, объемы которых соизмеримы с объемами аналогичных международных сетей 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 Интернет;

– 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.

Предложенная с учетом особенностей контента китайского сегмента Интернет модель информационного сервиса обеспечивает мониторинг актуальных новостных тем и сбор необходимого контента различной направленности.

В то же время, система мониторинга на основе использования каналов RSS является универсальной и может быть с минимальными перестройками адаптирована к любому национальному сегменту Интернет.

Литература

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. Ландэ Д.В., Березин Б.А., Додонов В.А. Обзор особенностей и возможности контент-мониторинга национального

сегмента сети Интернет // Реєстрація, зберігання і обробка даних, 2016. – Т. 18. – № 3. – С. 20-38.
7. Ландэ Д.В., Березин Б.А., Павленко О.Ю. Подход к мониторингу контента китайского сегмента Интернет // Міжнародна науково-практична конференція "Інтелектуальні технології лінгвістичного аналізу": Тези доповідей. – К.: НАУ, 2016. – С. 24.
8. 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.
9. 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.
10. 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.
11. 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.