УДК 004.891
DOI: 10.15827/0236-235X.116.085-088
Дата подачи статьи: 09.09.16
2016. Т. 29. № 4. С. 85–88

АРХИТЕКТУРА СИСТЕМЫ МОНИТОРИНГА ИНФОРМАЦИОННЫХ ТРЕНДОВ НА ОСНОВЕ СВОБОДНОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

С.А. Беляев, к.т.н., доцент, beliaev@nicetu.spb.ru;
А.В. Васильев, студент, unlike-2010@mail.ru
(Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), ул. Профессора Попова, 5, г. Санкт-Петербург, 197376, Россия);
С.А. Кудряков, д.т.н., зав. кафедрой, psi_center@mail.ru
(Санкт-Петербургский государственный университет гражданской авиации, ул. Пилотов, 38, г. Санкт-Петербург, 196210, Россия)

Статья посвящена описанию программной системы, предназначенной для выявления источников информационных трендов в результате анализа публикаций на новостных сайтах, в социальных сетях и блогах. Основная функция системы – построение графов распространения информации в сети Интернет.

Авторы обосновывают актуальность данной задачи, несмотря на наличие готовых решений, выполняющих сканирование данных в Интернет. Отдельно отмечена проблема лавинообразного увеличения объема информации, требующей обработки.

В статье описана модель для формализации процесса анализа информационных трендов и отмечены отличия от опубликованных подходов к решению данной задачи. Предложены основные шаги по автоматизации решения на основе данной модели. Особое внимание уделено возможности и обоснованности использования программных продуктов с открытым исходным кодом для решения отдельных подзадач. Для построения системы предлагается многоуровневая архитектура, демонстрирующая возможность рационального использования свободного ПО, дана последовательность работы системы.

На основе описанной архитектуры и предложенной модели разработан программный комплекс, обеспечивающий решение задачи мониторинга информационных трендов. Приведены результаты тестирования комплекса на примере нескольких новостных сайтов. Предложены подходы по дальнейшему развитию решения.

Ключевые слова: *информационные тренды, мониторинг Интернета, безопасность, система мониторинга, web-ресурсы, архитектура, модель.*

В современном мире информация начинает цениться гораздо выше природных ресурсов, потому что именно благодаря ей развиваются и разрушаются экономики, начинаются и заканчиваются войны, формируются и распадаются государства. Многие институты изучают распространение и влияние информационных потоков современного общества, которое все меньше читает книги и смотрит телевизор и все больше времени проводит в сети Интернет. Распространение деструктивной информации в Интернете может иметь катастрофические последствия, поэтому возможность контроля основных информационных трендов и источников их появления является важнейшей задачей. В частности, появляется все больше случаев судебных разбирательств, касающихся сообщений в социальной сети ВКонтакте [1].

Системы мониторинга публикаций в Интернете существуют много лет, активно используются в различных областях и сферах жизни и решают, помимо прочих, задачи анализа средств массовой информации [2], информации по рынку продаж, мероприятий, услуг, контроля появления сообщений на заданную тематику [3] и т.п. Большинство предлагаемых решений предоставляет ссылки на публикации и статистику упоминания тех или иных тем, но без построения графов распространения информации [4].

В Рунете существует множество новостных блогов и информационных сайтов, которые за сутки посещают миллионы пользователей. На наполнение большинства из них работает множество людей, обеспечивающих сбор, систематизацию и подготовку публикаций, зачастую сенсационные сообщения появляются из одного-двух источников. Использование автоматизированных средств мониторинга информационных трендов в системе МВД [5] позволит:

– на ранних стадиях выявлять, предупреждать, пресекать и раскрывать деятельность преступных групп, отдельных лиц и общественных объединений;

– выявлять тенденции по распространению информации;

– увеличить область охвата по сравнению с традиционными методами поиска;

– сократить время реагирования при возникновении новых информационных трендов.

Система мониторинга информационных трендов отслеживает любые изменения и появление новых вбросов информации на информационно-развлекательных порталах, страницах новостных интернет-изданий, блогах и прочих интернет-ресурсах, контролирует время размещения сообщений, предполагаемых авторов и формирует графы распространения.

Особенность публикаций в Интернете заключается в том, что распространением информации занимаются не только профессиональные авторы, зачастую в текстах встречаются опечатки или грубые ошибки, могут использоваться сокращения или выполнена замена слов на их синонимы, что затрудняет сопоставление текстов. Для ослабления влияния данных особенностей система мониторинга информационных трендов использует алгоритмы нечеткого поиска и словари синонимов и сокращений.

Основная трудность связана с объемом обрабатываемой информации: только в социальных сетях, используемых в России, количество сообщений увеличивается более чем на 500 миллионов сообщений в месяц [6], не говоря о новостных сайтах. Сообщения в социальных сетях зачастую имеют явный механизм цитирования, при использовании блогов и подобных им сайтов далеко не всегда можно найти ссылки на источники информации, а увеличение объема отдельных статей в зависимости от используемого алгоритма поиска дубликатов (плагиата) может иметь сложность более $O(n^2)$. Соответственно, поиск может осуществляться только выборочно и по ограниченному подмножеству поисковых запросов, так как построение системы, которая «знает все», равнозначно созданию специализированного аналога поисковых гигантов, таких как Google или Яндекс.

Для формализации процесса анализа информационных трендов целесообразно сформулировать задачу в виде математической модели: $M = (A, G, H, C, L, S, R, T)$, где $A = \{a\}$ – список интернет-адресов новостных сайтов, социальных сетей и блогов, подлежащих контролю, $G{:}A \rightarrow H$ – функция получения HTML-кода с сайтов с учетом внутренних переходов между страницами по ссылкам, здесь $H = \{h\}$ – множество полученного HTML-кода. $C{:}H \rightarrow L$ – функция предварительного анализа и формирования множества $L = \{l = <a, d, v>\}$, внутреннего представления публикаций, полученных из Интернета, в которых a – адрес публикации, d – дата ее появления, v – текст, подготовленный для анализа. $T = R(L, S)$ – множество обнаруженных информационных трендов, сформированных по элементам множества L с использованием словаря синонимов и сокращений $S=\{s\}$. Каждый элемент множества $T = \{t_i\}$ представляет собой кортеж $t_i = <d_i, gr_i, h_i>$, где d_i – дата и время появления информационного тренда; gr_i – граф появления информации в сети Интернет; h_i – HTML-код первоисточника информационного тренда. Граф появления информации в узлах содержит ссылки на публикации, по дугам осуществляется переход к информации о сайтах, на которых публикация появилась позже. В отличие от аналогов [7] предложенная модель не учитывает некоторые особенности информационных трендов, таких как определение цели, интенсивность воздействия, организация противодействия, что, с одной стороны, не позволяет в полной мере управлять информационными трендами, с другой – дает возможность упростить модель.

Ключевой особенностью предлагаемой модели является формализация информационных трендов в виде графов распространения информации в сети Интернет.

Модель функционирует следующим образом:

– сбор данных по заданному списку адресов и всем внутренним ссылкам;

– формирование внутреннего представления и предварительный анализ для дальнейшей обработки;

– построение графов появления информации в сети Интернет с использованием синонимов и алгоритмов нечеткого поиска.

Готовых решений, обеспечивающих построение графов появления информации в сети Интернет с учетом особенностей публикуемой информации, нет, но существуют программные продукты, в том числе свободное ПО, которые умеют частично решать данные задачи. Наиболее распространенные и универсальные – решения для локального хранения информации и выполнения нечеткого поиска, одно из наиболее популярных – Apache Lucene [8], в связи с этим дальнейшее построение системы выполнено с использованием данного продукта.

После получения HTML-документов из Интернета по заданному списку адресов осуществляется преобразование их во внутренний формат Lucene, а в качестве промежуточного представления используется JSON, создаваемый с помощью свободно распространяемой библиотеки GSON [9]. Библиотека GSON выбрана в связи с тем, что она быстрее формирует JSON для файлов объемом меньше 10 Мб по сравнению с JSON Simple, Jackson и JSONP. Решение на основе свободного ПО с открытым исходным кодом позволяет не только сократить время разработки, но и повысить его качество за счет использования отлаженных программных решений. Для полноценной реализации в системе мониторинга информационных трендов реализованы следующие алгоритмы:

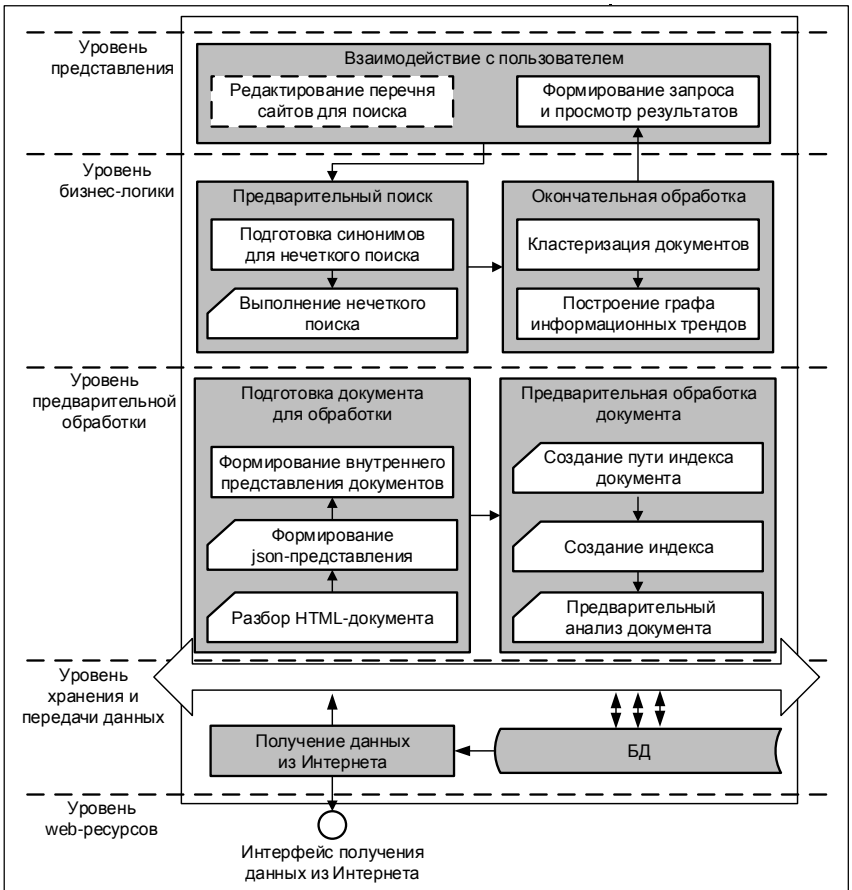
– анализ HTML-документа с целью выбора статей и обеспечения перехода по внутренним ссылкам;

– нечеткий поиск [10] с использованием синонимов и сокращений;

– статистический анализ степени совпадения двух текстов;

– определение последовательности появления информации в сети Интернет.

С учетом описанной математической модели, возможностей по использованию свободного ПО и требуемых к разработке алгоритмов спроектирована архитектура системы (см. рисунок). В архитектуре прямоугольниками со «срезанным» верх-



Архитектура системы

System architecture

ним левым углом представлены модули, реализованные с использованием свободного ПО.

Уровень web-ресурсов является источником данных, из которого с помощью интерфейса получения данных осуществляется сбор HTML-страниц. Полученные страницы поступают на уровень предварительной обработки, где сначала осуществляется подготовка документов, затем процесс обработки.

В рамках подготовки выполняются разбор HTML-страниц и выделение в них статей. В качестве инструмента выделения статей из страниц выбрана популярная библиотека с открытым исходным кодом Java HTML Parser [11]. Она обеспечивает API для извлечения данных и манипулирования ими, используя DOM, CSS и JQuery-подобные методы. Библиотека реализует WHATWG HTML5-спецификацию и разбирает HTML в ту же модель DOM, как это делают современные браузеры, такие как Google Chrome и Mozilla Firefox.

Затем формируется JSON-представление, а на его основании – внутреннее представление документа, которое может обрабатываться Apache Lucene. Дальнейшие этапы предварительной обработки по созданию пути индекса, созданию индекса и предварительному анализу документа выполняются непосредственно в Lucene. Результаты сохраняются в БД.

На уровне пользователя осуществляется настройка перечня сайтов для поиска, задается поисковый запрос и осуществляется визуальный анализ графов информационных трендов с просмотром соответствующих публикаций, их адресов и даты размещения.

Уровень бизнес-логики на основании запроса пользователя обеспечивает сбор необходимых синонимов для его уточнения, а затем нечеткий поиск с помощью механизмов Lucene. Результаты поиска кластеризуются с использованием стандартного алгоритма шинглов для определения меры совпадения двух документов, и для каждого кластера осуществляется построение графов распространения информации.

Тестирование системы проводилось на ограниченной выборке новостных сайтов (aif.ru, lenta.ru, ria.ru, rg.ru, bfm.ru, pikabu.ru). В результате было построено 25 графов распространения информации по этим сайтам. Среднее время выборки всех статей с одного сайта составило 24 минуты, среднее время выявления информационного тренда (обработка запроса, выделение кластера и построение графа распространения информации) – 27 секунд. В результате с одного сайта было обработано в среднем 230,65 Мб информации. Тестирование проводилось на персональном компьютере Lenovo y-500 с 6 Гб оперативной памяти и процессором Intel Core i5-3230 2.60GHz.

Оценка эффективности реализации показала, что значительно меньшую часть времени обработки запроса включает в себя нечеткий поиск с использованием синонимов и сокращений, что с учетом количества сайтов, использованных для тестирования, говорит о возможности дальнейшей оптимизации алгоритмов разбора HTML-страниц.

Использование данной системы позволяет осуществлять мониторинг появления информационных трендов и тем самым определять не только источник распространения конкретной новости, но и наиболее влиятельные источники информации, чтобы, с одной стороны, контролировать неожиданное появление новых источников, с другой – обеспечивать контроль минимального и достаточ-

ного количества информационных ресурсов. В дальнейшем целесообразно рассмотреть применимость различных вариантов алгоритмов, в первую очередь, разбора HTML-страниц и лишь затем кластеризации текстовой информации и возможности Lucene по настройке нечеткого поиска.

Литература

1. Борисов С.В., Васнецова А.С. Противодействие экстремистской деятельности – важный аспект обеспечения национальной безопасности // Правовая инициатива. 2014. № 3. URL: http://49e.ru/ru/2014/3/8 (дата обращения: 03.09.2016).
2. Мониторинг СМИ. URL: http://www.mlg.ru/solutions/pr/monitoring/ (дата обращения: 03.09.2016).
3. Программа СайтСпутник (FileForFiles & SiteSputnik) – поиск, сбор, мониторинг и анализ информации. URL: http://sitesputnik.ru/ (дата обращения: 03.09.2016).
4. SCAN Система комплексного анализа информации. URL: https://scan-interfax.ru/ (дата обращения: 03.09.2016).
5. Солодухин А.И., Романенко С.А., Беляев С.А., Медведева Я.И. Подход к построению комплексной системы предупреждения преднамеренных информационных трендов на основе семантического анализа текстовых ресурсов в сети Интернет // Актуальные проблемы психологической безопасности: сб. тр. регион. совещ. СПб: Свое Изд-во, 2012. С. 79–85.
6. Социальные сети в России: цифры и тренды за февраль 2016 года. URL: https://br-analytics.ru/blog/socialnye-seti-v-rossii-cifry-i-trendy-za-fevral-2016-g/ (дата обращения: 03.09.2016).
7. Потемкин А.В. Распознавание информационных операций средств массовой информации сети Интернет // Науковедение. 2015. Т. 7. № 2. URL: http://naukovedenie.ru/PDF/139TVN315.pdf (дата обращения: 03.09.2016). DOI: 10.15862/139TVN315.
8. Welcome to Apache Lucene. URL: http://lucene.apache.org/ (дата обращения: 03.09.2016).
9. Gson. URL: https://sites.google.com/site/gson (дата обращения: 03.09.2016).
10. Желудков А.В., Макаров Д.В., Фадеев П.В. Особенности алгоритмов нечеткого поиска // Инженерный вестн. 2014. № 12. URL: http://engsi.ru/file/out/745418 (дата обращения: 03.09.2016).
11. Jsoup: Java HTML Parser. URL: https://jsoup.org/ (дата обращения: 03.09.2016).

Software & Systems
DOI: 10.15827/0236-235X.116.085-088

Received 09.09.16
2016, vol. 29, no. 4, pp. 85–88

THE MONITORING OF INFORMATION TRENDS SYSTEM'S ARCHITECTURE BASED ON THE FREE SOFTWARE

S.A. Belyaev[1], *Ph.D. (Engineering), Associate Professor, beliaev@nicetu.spb.ru*
A.V. Vasilev[1], *Student, unlike-2010@mail.ru*
S.A. Kudryakov[2], *Dr.Sc. (Engineering), Head of Chair, psi_center@mail.ru*

[1] *St. Petersburg Electrotechnical University "LETI", prof. Popova St. 5, St. Petersburg, 197376, Russia*
[2] *St. Petersburg State University of Civil Aviation, Pilotov St. 38, St. Petersburg, 196210, Russian Federation*

Abstract. The article describes a software system designed to identify sources of information trends in the analysis of publications on news sites, social networks and blogs. The main feature of the system is the construction of a graph of information dissemination in the Internet.

The authors prove the relevance of this problem, despite the presence of ready-made solutions, which scan data in the Internet. The paper also focuses on a problem of exponential increase in the volume of information requiring processing.

The article describes a model to formalize the process of analyzing information trends and marked differences from published approaches to solving this problem. The authors propose basic steps for automation solutions based on this model. Special attention is paid to the possibility and validity of using software products with open source code for solving individual subtasks. To build the system the authors offer a layered architecture that demonstrates the possibility of rational use of free software and give a sequence of operation of the system.

On the basis of the architecture and the proposed model there is developed software that provides the solution to the problem of monitoring information trends. The results of testing based on several news sites. The paper proposes some approaches for further development of the solution.

Keywords: information trends, Internet monitoring, security, system monitoring, web-resources, architecture, model.

References

1. Borisov S.V., Vasnetsova A.S. Counteraction to the extremist activity – an important aspect of ensuring national security. *Pravovaya initsiativa* [The Legal Initiative]. Moscow, 2014, no. 3. Available at: http://49e.ru/ru/2014/3/8 (accessed September 3, 2016).
2. *Monitoring SMI* [Mass Media Monitoring]. Available at: http://www.mlg.ru/solutions/pr/monitoring/ (accessed September 3, 2016).
3. *Programma SaytSputnik (FileForFiles & SiteSputnik) – poisk, sbor, monitoring i analiz informatsii* [FileForFiles & SiteSputnik. Information Search, Capture, Monitoring and Analysis]. Available at: http://sitesputnik.ru/ (accessed September 3, 2016).
4. *SCAN Sistema kompleksnogo analiza informatsii* [SCAN. System Information Complex Analysis]. Available at: https://scan-interfax.ru/ (accessed September 3, 2016).
5. Solodukhin A.I., Romanenko S.A., Belyaev S.A., Medvedeva Ya.I. An approach to constructing a complex prevention system of malicious information trends based on text resource semantic analysis in the Internet. *Aktualnye problemy psikhologicheskoy bezopasnosti: sb. tr. region. soveshch.* [Proc. Regional Conf. on Important Problems of Psychological Security]. St. Petersburg, Svoe Izdatelstvo Publ., 2012, pp. 79–85 (in Russ.).
6. *Sotsialnye seti v Rossii: tsifry i trendy za fevral 2016 goda* [Social Networks in Russia: Numbers and Trends for February 2016]. Available at: https://br-analytics.ru/blog/socialnye-seti-v-rossii-cifry-i-trendy-za-fevral-2016-g/ (accessed September 3, 2016).
7. Potemkin A.V. Recognition of mass media information operations in the Internet. *Naukovedenie* [Science Studies]. 2015, vol. 7, no. 2 (27). Available at: http://naukovedenie.ru/PDF/139TVN315.pdf (accessed September 3, 2016).
8. *Welcome to Apache Lucene*. Available at: http://lucene.apache.org/ (accessed September 3, 2016).
9. *Gson*. Available at: https://sites.google.com/site/gson (accessed September 3, 2016).
10. Zheludkov A.V., Makarov D.V., Fadeev P.V. Characteristics of fuzzy search algorithms. *Inzhenerny vestnik* [Engineering Bulletin]. Moscow, N.E. Bauman MSTU Publ., 2014, no. 12. Available at: http:// engsi.ru/file/out/745418 (accessed September 3, 2016).
11. *Jsoup: Java HTML Parser*. Available at: https://jsoup.org/ (accessed September 3, 2016).