УДК 004.62

МОДЕЛИ ГЕНЕРАЦИИ КОНТЕНТА НОВОСТНЫХ ИСТОЧНИКОВ В СИСТЕМЕ МОНИТОРИНГА СОЦИАЛЬНЫХ МЕДИА-РЕСУРСОВ

MODELS GENERATING THE CONTENT OF NEWS SOURCES IN A SISTEM OF MONITORING SOCIAL MEDIA RESOURCES

А.А. Овсянников, А.А. Смирнов
A.A. Ovsyannikov, A.A. Smirnov

Федеральное государственное казённое военное образовательное учреждение высшего образования «Академия Федеральной службы охраны Российской Федерации», Россия, 302034, Орёл, ул. Приборостроительная, 35

Federal state military educational institution of higher professional education "Academy of the Federal security service of the Russian Federation", 35 Priborostroitelnaya St, Orel, 302034, Russia

ovsyannikov.aa@mail.ru, al2smi@gmail.com

Аннотация. В статье рассматривается формирование и анализ моделей генерации контента новостными источниками для решения задач сбора и обработки сообщений в системе мониторинга и анализа данных. Выделены особенности форм генерации сообщений, позволяющие обосновать их разделение на классы. Предложены несколько классов источников с разбиением по частоте и неравномерности их публикаций. Учет специфики моделей генерации позволяет повысить эффективность системы сбора данных в условиях ограничения технических ресурсов (мощности процессора, пропускной способности канала и др.) Наличие в системе сбора обратной связи с источником генерации контента позволяет учесть факторы связанные с неравномерностью публикаций и ограничениями вычислительных ресурсов.

Resume. The article deals with the formation and analysis of models of the content of news sources for communications solutions for collecting and processing tasks in the system of monitoring and data analysis. The features of the forms generate reports that allow them to justify the division into classes. Proposed several classes of sources with the division in frequency and irregularity of their publications. Accounting for the specifics of the generation of models to improve the efficiency of data collection system in conditions of limited technical resources (processing power, bandwidth, etc.). The presence of the system of collecting feedback from the source content generation allows to take into account factors related to the non-uniformity of publications and limited computing resources.

Ключевые слова: Система мониторинга, сбор данных, большие данные, обработка больших объемов данных, интеллектуальный анализ данных, аналитическая обработка данных, адаптивные системы.

Keywords: Monitoring system, data collection, big data, processing large amounts of data, data mining, analytical data processing, adaptive systems.

Введение

На сегодняшний день высокую популярность имеют социальные медиа-ресурсы, в частности социальные сети [Mika, 2007], которые позволяют не только публиковать новостную информацию, но и обсуждать события, описываемые в новостях. Но продолжающиеся изменения в интернет-ресурсах приводят, как правило, к сложностям их автоматической обработки и вызывают необходимость получения и применения в моделях мониторинговых систем результатов наблюдений за влиянием внешней среды. В связи с этим большое количество источников информации, неравномерность потоков текстовых сообщений, отсутствие обратной связи и дефицит информации для управления процессами затрудняют получение сообщений, и в целом – снижают качество результатов мониторинга, что требует развития методов сбора и обработки новостной информации социальных медиа. С теоретических позиций гносеологии необходимо искать новые формы и методы познания, рефлексии развивающейся распределенной среды. С общих позиций методологии это требует разработки подходов регистрации и анализа данных о влиянии латентных факторов на значимые свойства новостных интернет-ресурсов.

Развитие современных методов обработки больших объемов данных, практические потребности повышения «интеллектуальности» процедур и средств обработки, учета закономерностей в данных и процессах, обусловливают необходимость развития научно-обоснованных спо-

собов сбора данных, учитывающих изменчивость (и слабую предсказуемость) поведения источников и определяют актуальность изучения особенностей моделей генерации новостных сообщений.

Автоматический сбор данных в сети Интернет осуществляют программы называемые поисковыми роботами или краулерами [Pant et al., 2004, Cristopher, Mark, 2010]. Отдельными группами исследователей ведутся работы направленные на повышение эффективности сбора и обработки данных. Рабачевский Е.А., Цукерман А.Н., Иванова О.В., Иванов П.В., Смелов М.Н., Коляда А.С., Гогунский В.Д. в своих работах уделяют внимание повышению эффективности сбора информации путем совершенствования организационных составляющих процесса сбора [Рабачевский, Цукерман, 2016, Иванова и др., 2010, Коляда, Гогунский, 2014]. Аюков С.В., Бартунов О.С., Родичев Е.Б., Печников А.А., Чернобровкин Д.И. исследуют механизмы оптимизации и адаптации сбора информации [Аюков и др.,2002, Печников, Чернобровкин, 2012]. В связи с этим речь может идти о моделях взаимовлияния (взаимосвязи) исследуемых процессов [Жиляков, Белов, 2014].

Несмотря на повышенное внимание к мониторингу информации сети Интернет, по мнению авторов, недостаточное внимание уделяется исследованию специфики взаимодействия средств сбора данных с интернет-ресурсами, а также обеспечению эффективности сбора сообщений в интернет-СМИ, блогосфере, социальной сети и контролю качества результатов. Далее в статье рассматриваются способы идентификации и построения моделей генерации новостных сообщений.

Этап идентификации и построения моделей генерации новостных сообщений нацелен на уменьшение избыточных действий по сбору новостных сообщений, приводящих к расходу ресурсов в системе мониторинга разнородных потоков новостей. Для создания классификационных группировок видов потоков применен способ исследования источников на основе базы прецедентов.

Для этого осуществлялось формирование информационной базы исследования, выполнен сбор сообщений трех крупных социальных сетей: Twitter, ВКонтакте, Google+. Загрузка сообщений проводилась с использованием загрузчика (краулера) собственной разработки. Сообщения загружались в базу данных (БД) с регистрацией информации о дате и времени публикации сообщения, дате и времени загрузки сообщения в БД, сведений об авторе и др.

Использование методов анализа данных и процессов позволило сформировать показатели, характеризующие частоту публикации сообщений и неравномерность их публикации. С использованием кластерного анализа сообщения были разбиты на три группы по каждому из показателей. В совокупности три группы по двум показателям описывали девять возможных классов источников.

Изучена работа групп источников, наиболее сходные по работе группы объединены алгоритмом кластерного анализа [Барсегян и др., 2009]. После объединения классов были идентифицированы четыре типовые модели генерации контента публикаций.

Поэтому для идентификации и построения моделей генерации контента были осуществлены следующие действия:

1. Осуществлен сбор сообщений для формирования массива данных (база прецедентов).
2. Известными методами анализа данных [Барсегян и др., 2009] выявлены характеристики и рассчитаны показатели, характеризующие отклонения в поведении источника публикаций.
3. С использованием набора показателей сформированы классы источников с учетом характерных особенностей генерации контента.
4. Выполнено объединение классов в несколько обобщенных моделей генерации контента.

В процессе анализа базы прецедентов выявляются качественно-количественные характеристики, отражающие особенности генерации контента источниками публикации новостных сообщений.

Обозначим за I – источник сообщений. В качестве источника сообщений в социальной сети выступает пользователь или социальная группа пользователей, которые могут создавать сообщения, другими словами, генерировать контент.

Характеристики извлекаемые в процессе анализа источника:

C_I– количество сообщений источника загруженных в систему анализа за установленный временной промежуток;

D_I – количество дней анализа источника, рассчитываем как разницу текущей даты и даты начала анализа.

На основе данных характеристик рассчитываем:

F_I – частота публикации источника, рассчитываемая как отношение количества сообщений к периоду публикации в днях.

$$F_I = \frac{C_I}{D_I} \quad (1)$$

U_I – неравномерность публикации сообщений источником, рассчитывается как отклонение от среднего числа публикаций в день.

$M_{I,d}$ – количество сообщений за день d для I-го источника

$$U_I = \frac{1}{C_I}\sum_{D_I}(F_I - M_{I,d})^2 \quad (2)$$

Для построения и классификации источников по рассчитанным показателям было осуществлено нормирование F_1 и U_1.

Для проведения «полевого» исследования был создан программный комплекс для извлечения, обработки и анализа характеристик источников. В системе имеются модули, позволяющие извлекать данные из крупных социальных сетей (Twitter, Вконтакте, Google+) посредством API функционала предоставляемого социальными сетями. Каждый модуль загрузки сообщений имеет средства анализа характеристик процесса загрузки информации. Для хранения данных используется реляционная база данных MSSQL 2008.

В ходе эксперимента были обработаны данные от 5380 источников сообщений. Получены данные по частоте публикации источников в течение месяца F_1 и данные о равномерности публикации U_1 сообщений источником.

Для построения классификации (классификационной группировки) по типам источника было выполнено разбиение частоты публикации на три класса-диапазона (высокая, средняя, низкая), равномерность публикации так же разбита на три класса (высокая, средняя, низкая).

Комбинация классов частоты и неравномерности публикации дает девять классов типов источника (см. табл.)

Таблица
Table

Символьное обозначение классов типов источников
Symbols classes types of sources

Неравномерность / Частота	Высокая	Средняя	Низкая
Высокая	ВВ	ВС	ВН
Средняя	СВ	СС	СН
Низкая	НВ	НС	НН

В отдельных классах сообщений было мало или они вообще отсутствовали, эти классы в дальнейшем нами не использовались. Оставшиеся классы были объединены в четыре группы, которые характеризуют основные модели генерации контента.

В процессе исследований социальных медиа были изучены схемы функционирования новостных источников, что позволило сформировать четыре основных модели генерации контента новостными источниками информации (см. рис. 1):

- Публикации с фиксированным интервалом между сообщениями (F~0-1, U~0-0.3) – публикации сообщений через определенные временные интервалы (рис. 2а). Время между публикациями постоянно (неравномерность низкая), частота публикации различна (высокая, средняя или низкая).

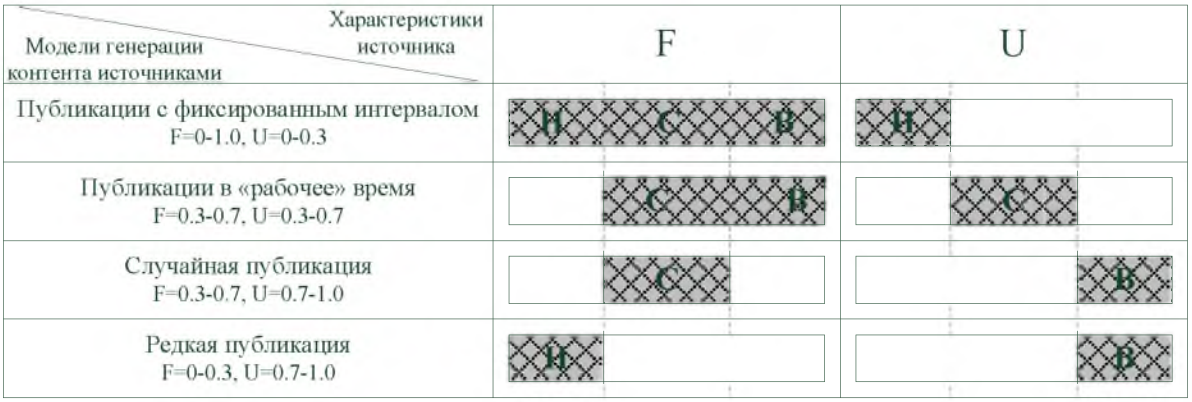

Рис. 1. Распределение параметров по моделям генерации контента
Fig. 1. Distribution of options by model content generation

- Публикации в «рабочее» время (F~0.3-1.0, U~0.3-0.7) – публикация материалов в определенные временные промежутки (рис. 2б), например, утром и после обеда. Время между публикациями различно, частота высокая или средняя.

- Случайная публикация (F~0.3-0.7, U~0.7-1.0) – публикация в различное время в течение дня (рис. 2в). Время между публикациями различно, частота средняя.

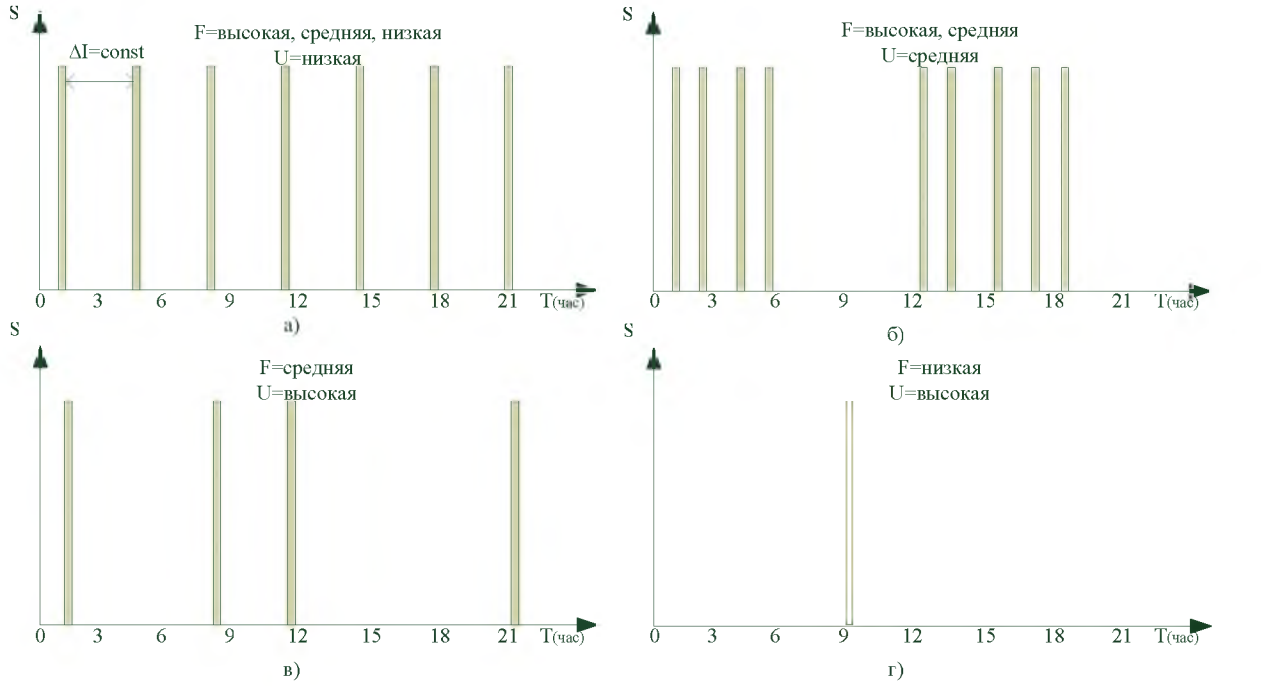

Рис. 2. Следование сообщений в различных моделях генерации контента
Fig. 2. Following reports in various models of content generation

- Редкая публикация (F~0-0.3, U~0.7-1.0) – непостоянные публикации, например раз в несколько дней (рис. 2г). Время между публикациями различно, частота низкая.

Практические эксперименты показали, что возможны комбинации представленных моделей или замещение одной модели другой на некоторый промежуток времени.

В работе [Аюков и др., 2002] каждому документу сопоставляют величину характеризующую изменчивость документов во времени применительно к веб-страницам и в соответствии с этой величиной определяют период обхода документов. В работе [Печников, Чернобровкин, 2012] описан адаптивный краулер для эффективного сбора гиперссылок с различных Web-ресурсов. Предложенные в статье способы идентификации и формирования моделей генерации являются дополнением к известным подходам использования адаптивных алгоритмов для построения систем интернет-мониторинга.

В настоящее время исследование особенностей новостных потоков и эксперименты по изучению факторов, влияющих на изменения моделей генерации продолжаются. Их итоги и алгоритмы интернет-мониторинга являются предметом рассмотрения следующей публикации.

Заключение

Для осуществления эффективного собора сообщений необходимо правильно представлять работу источников сообщений, проводить исследование методик публикации сообщений, что позволит сформировать модели работы новостных источников. Типизация форм поведения источников позволяет упростить получение информации из источников, уменьшить время поиска, классификации и обработки сообщений. В целом использование знаний о поведении источников позволяет экономить ресурсы и осуществить построение эффективных систем сбора новостных сообщений, использующих адаптивные алгоритмы.

Разработанные способ формирования и модели генерации возможно использовать при решении задач информационно-аналитической поддержки деятельности информационных служб государственного и коммерческого сектора.

Список литературы
References

Аюков С.В., Бартунов О.С., Родичев Е.Б. 2002. Оптимизация сканирования ресурсов Интернет поисковой машиной с помощью оценок скорости изменения документов. Научный сервис в сети Интернет: сб. Материалов всероссийской научн. Конф., Новороссийск, 23-28 сентября 2002 г. МГУ им. М.В. Ломоносова. 133-137.

Ayukov S.V., Bartunov O.S., Rodichev E.B. 2002. Internet resources scan search engine optimization with the help of documents change rate estimates. Scientific service on the Internet: Sat. Russia scientific materials. Conf., Novorossiysk, 23-28 September 2002. MSU. MV Lomonosov: 133-137.

Барсегян А.А., Куприянов М. С., Холод И.И., Тесс М.Д., Елизаров С.И. 32009. Анализ данных и процессов: учеб. Пособие. 3-е изд., перераб. и доп. СпБ.: БХВ-Петербург. 512.

Barseghyan A.A., Kupriyanov M.S., Frost I.I., Tess M.D., Elizarov S.I. 2009. Analysis of the data and processes: Proc. Manual. 3rd ed., Revised. And ext. spb., BHV-Petersburg. 512.

Жиляков Е.Г., Белов С.П. 2014. Об оценивании параметров линейных моделей многомерных сигналов. Научные ведомости БелГУ. Сер. История. Политология. Экономика. Информатика. 8(179): 83-89

Zhilyakov E.G., Belov S.P. 2014. Ob ocenivanii parametrov linejnyh modelej mnogomernyh signalov. Nauchnye vedomosti BelGU. Ser. Istorija. Politologija. Jekonomika. Informatika. [Parameter Estimation of linear models of multidimensional signals. Belgorod State University Scientific Bulletin. History Political science Economics Information technologies] 8(179): 83-89.

Иванова О.В., Иванов П.В., Смелов М. Н. 2010. Проблемы и алгоритмы поиска информации в глобальных компьютерных сетях. T-Comm. 3: 23-25

Ivanova O.V., Ivanov P.V., Smelov M.N. 2010. Problems and algorithms for finding information on global computer networks. T-Comm. 3: 23-25

Коляда А. С., Гогунский В. Д. 2014. Извлечение информации из слабоструктурированных вебстраниц. ВЕЖПТ. 9 (67)

Kolyada A.S., Gogunsky V.D. 2014. Extracting information from semi-structured web pages. VEZHPT. 9 (67)

Печников А.А., Чернобровкин Д.И. 2012. Адаптивный краулер для поиска и сбора внешних гиперссылок. Управление большими системами. 36: 301-315

Pechnikov A.A., Chernobrovkin D.I. 2012. Adaptive crawler to search for and collect external hyperlinks. Managing large systems. 36: 301-315

Рабачевский Е.А., Цукерман А.Н. 2013. Некоторые аспекты задачи исследования распространения информации в социальной сети вконтакте [Электронный ресурс] Национальный Открытый Университет «ИНТУИТ». Режим доступа: http://rabchevsky.name/sites/default/files/download/aist13.pdf (10 июня 2016).

Rabachevsky E.A., Zuckerman A.N. 2013. Some aspects of the problem of research dissemination of information in the social network vkontakte [Electronic resource] The National Open University "INTUIT" Access: http://rabchevsky.name/sites/default/files/download/aist13.pdf (Accessed 10 June 2016)

Cristopher O., Mark N. 2010. Web Crawling. Foundations and Trends in Information Retrieval, Volume 4, Issue 3, March 2010: 175-246

Mika P. 2007. Social Networks and the Semantic Web (Semantic Web and Beyond), Springer, 234.

Pant G., Srinivasan P., Menczer F. 2004. Crawling the Web. In “Web Dynamics”. M. Levene, Poulovassilis, eds. Springer: 153–178.