УДК 519.68
DOI: 10.15827/0236-235X.121.095-098
Дата подачи статьи: 26.12.17
2018. Т. 31. № 1. С. 095–098

МОДЕЛЬ ОТКРЫТОГО КУБА ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ В СОЦИАЛЬНЫХ СЕТЯХ

А.В. Иващенко [1], *д.т.н., профессор, anton.ivashenko@gmail.com*
Н.М. Шлычкова [1], *студентка, kler7409@yandex.ru*
В.А. Исайко [2], *инженер, visayko@gmail.com*
П.В. Ситников [2], *к.т.н., директор, sitnika@o-code.ru*

[1] *Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе, 34, г. Самара, 443086, Россия*
[2] *ООО «Открытый код», ул. Ярмарочная, 55, г. Самара, 443001, Россия*

Социальные сети можно рассматривать как важнейший источник больших данных, описывающих взаимодействие пользователей в процессе обмена информацией. Выявление закономерностей на этом уровне позволяет определять основные особенности поведения, выявлять информационное влияние, обнаруживать и анализировать колебания интересов пользователя к различным информационным объектам и событиям и т.п. Применение такого рода аналитики на практике позволяет реализовывать эффективную контекстную рекламу, повышать эффективность работы социальных сетей, а также решать различные проблемы информационной безопасности.

Анализ больших данных, описывающих взаимодействие пользователей социальных сетей, является сложной технической проблемой: необходимо интегрироваться с несколькими социальными сетями для импорта данных, ассоциировать отдельные профили одних и тех же пользователей в разных сетях, сопоставлять факты их взаимодействия с реальными событиями и выявлять основные тенденции и отклонения.

Для решения этой задачи предлагается модель открытого куба, основанная на построении ортогональной системы индикаторов, характеризующих изменение данных в зависимости от разных факторов. При этом производится распределение возникающих событий взаимодействия относительно пользователей, динамики развития их интереса во времени, реакции на внешние события и т.д. с помощью инструментария взаимного корреляционного анализа временных рядов с использованием интервальных корреляционных функций.

В данной статье описываются основные проблемы анализа больших данных в социальных сетях, предлагаемая модель открытого куба и алгоритм анализа данных, позволяющий выявлять отклонения в поведении пользователей социальных сетей.

Описанная модель и ее реализация были испытаны и апробированы с использованием типового набора данных, полученных из ряда социальных сетей. В дополнение к реальному регулярному набору результатов переговоров пользователей социальных сетей была введена партия сообщений, генерируемых онлайн-ботом, существование которого было выявлено посредством предложенного подхода.

Ключевые слова: *социальные сети, большие данные, анализ, открытый куб.*

Существует достаточно много различных источников данных для анализа поведения пользователей сети Интернет. Например, социальные сети, новостные порталы, ленты, где люди могут под разными аватарами давать информацию любого рода. Такая информация носит заведомо субъективный характер, что необходимо учитывать при ее анализе. Многие открытые источники информации, онлайн-энциклопедии и средства массовой информации стараются решить эту проблему путем реализации массового обсуждения информационного контента, определения политик рецензирования и модерации, внедрения систем рейтингов, взаимной оценки и т.п. Однако вопрос оценки объективности информации, публикуемой в этих источниках, остается открытым.

Решить данную проблему могут аналитические инструменты выявления интереса пользователей сети Интернет на основе анализа их поведения, отраженного в различных информационных источниках. Для этого необходимо семантический анализ публикуемого контента дополнить анализом потоков событий, характеризующих его создание, прочтение, обсуждение и изменение. Учитывая большой объем такого рода событий, их высокое многообразие и изменчивость, а также слабую структурированность, при реализации аналитических инструментов необходимо использовать технологии анализа больших данных [1].

Исследованию фундаментальных принципов функционирования социальных сетей, информационному влиянию и управлению социально-экономическими системами с их использованием в настоящее время уделяется достаточно существенное внимание [2–4]. Описанию трендов развития социальных сетей и возможностям автоматизированного анализа данных посвящены работы [5, 6]. Моделирование виртуальных сообществ и выявление интереса пользователей для последующего анализа их развития и построения эффективных функциональных инструментов позволили реализовать ряд полезных проектов в этой области [7, 8].

Однако современные тенденции в развитии Интернета [9] потребовали реализации новых теоретических подходов.

На практике в настоящее время наиболее развиты системы анализа социальных сетей для коммерческих организаций. Независимо от того, для кого разрабатываются такие системы, их можно классифицировать по следующим пунктам: уровни анализа, модели, объекты анализа открытых источников сети Интернет, методы анализа, режимы анализа и сбора, охват источников и объем обрабатываемых данных. Система подобного рода может использоваться как для решения задач внутри самой организации, так и за ее пределами. На данный момент на рынке наиболее развитыми являются системы, направленные прежде всего на управление взаимоотношениями с клиентами. В целом существующие системы могут предоставлять следующие возможности: мониторинг упоминания брендов, определение рыночных рисков и возможностей, веб-аналитика, поддержка работы в онлайновых социальных сетях, прогнозирование и управление социальными сетями. В частности, система Radian 6 предназначена для отслеживания в реальном времени упоминаний брендов с учетом тональности в социальных сетях и для участия в происходящих обсуждениях. Система Alterian SM2 позволяет отслеживать упоминания брендов в социальных сетях с учетом тональности: положительная, отрицательная, нейтральная. Кроме того, данная система позволяет локализовать места обсуждений и определять демографические характеристики пользователей социальных сетей. Система анализа социальных сетей BrandSpotter позиционируется как система мониторинга и управления репутацией бренда в социальных сетях, а также упоминания бренда с учетом тональности; отслеживаются наиболее значимые пользователи социальных сетей по данной тематике.

Для автоматизации анализа поведения пользователей социальных сетей требуется система, которая могла бы проводить мониторинг подобных изменений. Мониторинг как таковой включает в себя получение и структурирование первичных данных. Собираются такие данные, как тексты сообщений, опубликованные материалы, ссылки на внешние ресурсы и прочее. Возможности системы во многом зависят от используемых данных и от способа их обработки. Анализ подразумевает несколько этапов обработки первичных данных, таких как вычисление базовых показателей и выявление статистических и структурных закономерностей, дающих понимание природы исследуемой сети. Прогноз возможен после идентификации математической модели информационного процесса. Могут использоваться статистические модели и модели динамических процессов на графах, семантических сетях и т.п. Управление заключается в оказании целенаправленных воздействий на социальную сеть для перевода информационных процессов в желаемое состояние. Задачи по анализу, прогнозированию и управлению могут быть разными, в первую очередь, в зависимости от того, кто ее ставит, то есть кто является конечным пользователем системы.

Для решения этой задачи предлагается технология открытого куба, основанная на построении ортогональной системы индикаторов, характеризующих изменение данных в зависимости от разных факторов. При этом производится распределение возникающих событий взаимодействия относительно пользователей, динамики развития их интереса во времени, реакции на внешние события и т.д. с использованием инструментария взаимного корреляционного анализа временных рядов с использованием интервальных корреляционных функций [10].

Представим потоки событий информационной активности и взаимодействия пользователей открытых ресурсов Интернета в виде булевых переменных:

$$e_{i,j,k} = e_{i,j,k}(u_i, w_j, t_{i,j,k}) = \{0,1\}, \quad (1)$$

где u_i – пользователь (актор); w_j – информационный объект (статья, пост или комментарий); $t_{i,j,k}$ – время внесения изменений.

Логическую функцию, определяющую отношение события к выбранному индикатору, определим в виде

$$d_n\left(e_{i,j,k}, \delta_n\left(e_{i,j,k}\right), \Delta t_n^{(f)}\right) = \\ = \begin{cases} 1, & \delta_n\left(e_{i,j,k}\right) \wedge \left(t_{i,j,k} \in \Delta t_n^{(f)}\right), \\ 0 & \text{иначе}, \end{cases} \quad (2)$$

где $\Delta t_n^{(f)} = \left(t_{\text{н}}^{(f)}, t_{\text{к}}^{(f)}\right)$, а $\delta_n\left(e_{i,j,k}\right)$ – условие отнесения события к выбранному индикатору.

Наличие линейной связи между потоками событий $\{e_{i1,j1,k1}\}$ и $\{e_{i2,j2,k2}\}$ в разложении по выбранным индикаторам d_1, d_2 соответственно в этом случае можно представить в виде

$$K_J\left(\left\{e_{i1,j1,k1}\right\}, \left\{e_{i2,j2,k2}\right\}, d_1, d_2\right) = \\ = \sum_M \sum_{i1,j1,k1} \sum_{i2,j2,k2} d_1\left(e_{i1,j1,k1}, \delta_1\left(e_{i1,j1,k1}\right), J\Delta\tau\right) \times \\ \times d_2\left(e_{i2,j2,k2}, \delta_2\left(e_{i2,j2,k2}\right), (J+M)\Delta\tau\right). \quad (3)$$

Совокупность $\{d_n\}$ назовем открытым кубом.

На основе предоставленной модели разработан алгоритм для анализа больших данных в социальных сетях, который состоит из двух этапов.

Этап 1. Расчет вектора частоты выборки для всех пользователей и разработка стандартного вектора отклонения для различных пользователей.

Необходимо для набора событий $\{e_{i,j,k}\}$ сформировать показатель:

$$\left(\Omega_m, \frac{1}{\sum_{i,j,k} d_1\left(e_{i,j,k}, \left[w_j \in \Omega_m\right], \Delta t_1^{(\text{пуб})}\right)}\right), \quad (4)$$

где $\Delta t_1^{(\text{пуб})}$ – время (интервал) публикации; $\lfloor w_j \in \Omega_m \rfloor$ – условие соответствия информационных объектов (постов) тематике Ω_m, а знаменатель содержит суммарное число пользователей, опубликовавших схожие посты по данной теме за период $\Delta t_1^{(\text{пуб})}$.

Для данного показателя необходимо также определить СКО $\sigma_{i,j,k}^m$.

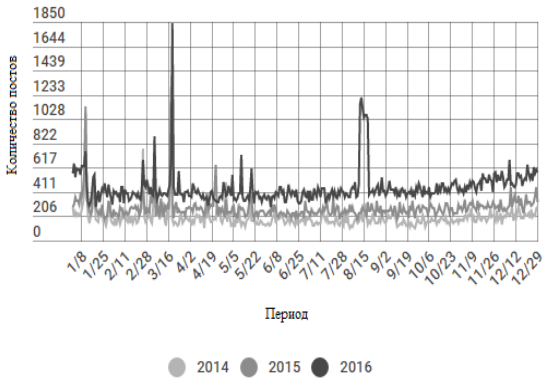
Этап 2. Вычисление показателя отклонения для конкретного пользователя. На данном этапе необходимо выбрать темы Ω_m и преобразовать их в представление ключ-значение, после чего обработать пары данных и подсчитать сумму тематик с одним и тем же ключом.

Для реализации предлагаемого подхода было разработано программное решение для идентификации фокуса в социальных сетях на основе обнаружения и анализа больших данных.

Решение может интегрироваться с различными источниками данных, идентифицировать тематики в виде облаков тегов и обрабатывать их изменения во времени. Данные, импортированные из социальных сетей, фиксируются в БД и могут обрабатываться либо в режиме реального времени, либо в пакетном режиме. Краулер обращается асинхронно к web-сервису с запросами на получение данных из социальных сетей. Получив запрос, web-сервис подтверждает начало обработки данного запроса. Далее web-сервис обращается к интегратору, который начинает выгружать запрошенные данные в виде RDF/XML-файлов, сохраняя промежуточные выгруженные данные, чтобы единым блоком передать уже выгруженные. Далее в фоновом режиме, то есть в режиме, при котором нет необходимости контролировать процесс выгрузки данных, интегратор автоматически продолжает ранее запущенный процесс, грузит данные в БД и с помощью Apache JENA формирует RDF/XML-файлы для последующей передачи.

Описанная модель, программное решение и его реализация были испытаны и апробированы с использованием типового набора данных, полученных из ряда социальных сетей. В дополнение к реальному регулярному набору результатов переговоров пользователей социальных медиа была введена партия сообщений, генерируемых онлайн-ботом. Помимо социальных медиа (без предварительного знания о структуре данных), алгоритмы анализа больших данных смогли выявить влияние онлайн-бота.

Результаты представлены на рисунке, где продемонстрированы ежегодные тенденции активности пользователей. Пик, определенный 15 августа, соответствует активности бота и может быть легко определен агентом, сравнивающим поведение предыдущих периодов. Описанные результаты исследований показывают, что предлагаемую модель



Анализ активности пользователей социальной сети

Analysis of social network users' activity

можно использовать для анализа поведения в сети и выявления негативного информационного влияния.

Таким образом, предлагаемая модель позволяет фиксировать процесс деятельности пользователя Интернета с учетом сочетания человеческого и временного факторов. Выявление закономерностей позволяет определять основные особенности поведения, информационное влияние, устанавливать и анализировать колебания интересов пользователя к различным информационным объектам и событиям и т.п. Применение такого рода аналитики на практике позволяет реализовывать эффективную контекстную рекламу, повышать эффективность работы социальных сетей, а также решать различные проблемы информационной безопасности.

Литература

1. Bessis N., Dobre C. Big data and internet of things: a roadmap for smart environments. Berlin, Springer, 2014, 450 p.
2. Базенков Н.И., Губанов Д.А. Обзор информационных систем анализа социальных сетей // Управление большими системами. 2013. Вып. 41. С. 357–394.
3. Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Социальные сети: модели информационного влияния, управления и противоборства. М.: Физматлит, 2010. 228 с.
4. Бреер В.В., Новиков Д.А., Рогаткин А.Д. Управление толпой: математические модели порогового коллективного поведения. М.: ЛЕНАНД, 2016. 168 с.
5. Wei W., Joseph K., Liu H., Carley K. Exploring characteristics of suspended users and network stability on Twitter. Social Network Analysis and Mining, 2016, pp. 6–51.
6. Kadushin C. Understanding social networks: theories, concepts, and findings. Oxford Univ. Press, 2012, 264 p.
7. Орлов А.Ю., Иващенко А.В. Организация виртуального сообщества в сети Интернет // Информационные технологии. 2008. № 8. С. 15–19.
8. Иващенко А.В., Пугачева Е.С., Погодина С.С. Моделирование виртуальных сообществ пользователей интегрированной информационной среды // Управление большими системами: сб. тр. М.: Изд-во ИПУ РАН, 2010. Вып. 29. С. 68–87.
9. One Internet. Global commission on Internet Governance. Ghatham House: The Royal Institute of International Affairs, 2016. URL: https://www.cigionline.org/initiatives/global-commission-internet-governance (дата обращения: 01.11.2017)
10. Прикладной анализ случайных процессов; [под ред. С.А. Прохорова]. Самара: Изд-во СамНЦ РАН, 2007. 582 с.

Software & Systems
DOI: 10.15827/0236-235X.121.095-098

Received 26.12.17
2018, vol. 31, no. 1, pp. 095–098

THE OPEN CUBE CONCEPT FOR ANALYSING SOCIAL MEDIA BIG DATA

A.V. Ivaschenko [1], *Dr.Sc. (Engineering), Professor, anton.ivashenko@gmail.com*
N.M. Shlychkova [1], *Student, kler7409@yandex.ru*
V.A. Isayko [2], *Engineer, visayko@gmail.com*
P.V. Sitnikov [2], *Ph.D. (Engineering), Director, sitnika@o-code.ru*

[1] *Samara National Research University, Moskovskoe Highway 34, Samara, 443086, Russian Federation*
[2] *"Otkryty kod" LLC, Yarmarochnaya St. 55, 443001, Samara, Russian Federation*

Abstract. Modern social media can be treated as an important source of Big Data describing users' behavior during informational exchange. Understanding the basic trends at this level can help to determine the main behavior features, identify informational influence and analyze deviations of users' interest to various informational objects and events, etc. On a practical level, this kind of analytics allows implementing an effective context-based advertising, increasing the efficiency of social networks functionality and solving various problems of information security.

Analysis of Big Data that characterize social media users' interaction appear to become a complex technical problem. The reasons are: it is required to integrate with several social networks for data import, to associate independent profiles of the same users at different networks, to correlate the facts of their interaction with real events and determine basic trends and deviations.

To solve the problem the authors propose to implement a technology of "open cube" based on an orthogonal indicators system describing the data change dynamics in time depending on different factors. It is proposed to analyze distribution of incoming user interaction events, dynamics of their interest evolution in time, reaction to incoming events, etc. using cross-correlation analysis of time series using interval-based functions.

The paper describes the basic problems of Big Data analysis in social media, the proposed abstract model of the open cube and the data analysis algorithm that allows identification of users' activity at social media. The described model and its implementation were tested using a typical data set derived from a number of social networks. In addition to a real regular data set of social media users' negotiation there was also introduced a series of messages generated by a bot, which was successfully identified using the proposed approach.

Keywords: social media, big data, analysis, open cube.

References

1. Bessis N., Dobre C. *Big data and internet of things: a roadmap for smart environments*. Berlin, Springer Publ., 2014, 450 p.

2. Bazenkov N.I., Gubanov D.A. Information systems for social networks analysis: a survey. *Upravlenie bolshimi sistemami* [Large-Scale Systems Control]. 2013, no. 41, pp. 357–394 (in Russ.).

3. Gubanov D.A., Novikov D.A., Chkhartishvili A.G. *Sotsialnye seti: modeli informatsionnogo vliyaniya, upravleniya i protivoborstva* [Social networks: models of information influence, governance and confrontation]. Moscow, Fizmatlit Publ., 2010, 228 p.

4. Breer V.V., Novikov D.A., Rogatkin A.D. *Upravlenie tolpoy: matematicheskie modeli porogovogo kollektivnogo povedeniya* [Crowd management: mathematical models of threshold collective behavior]. Moscow, LENAND Publ., 2016, 168 p.

5. Wei W., Joseph K., Liu H., Carley K. Exploring characteristics of suspended users and network stability on Twitter. *Social Network Analysis and Mining*. 2016, pp. 6–51.

6. Kadushin C. *Understanding social networks: theories, concepts, and findings*. Oxford Univ. Press, 2012, 264 p.

7. Orlov A.Yu., Ivashchenko A.V. Organization of a virtual community on the Internet. *Informatsionnye tekhnologii* [Information Technologies]. 2008, no. 8, pp. 15–19 (in Russ.).

8. Ivashchenko A.V., Pugacheva E.S., Pogodina S.S. Modeling virtual communities of users of the integrated information environment. *Upravlenie bolshimi sistemami* [Large-scale systems control]. Moscow, ISU RAS Publ., 2010, no. 29, pp. 68–87 (in Russ.).

9. *One Internet. Global commission on Internet Governance*. Ghatham House, The Royal Institute of Int. Affairs, 2016. Available at: https://www.cigionline.org/initiatives/global-commission-internet-governance (accessed November 1, 2017).

10. *Prikladnoy analiz sluchaynykh protsessov* [Applied analysis of random processes]. S.A. Prokhorov (Ed.). Samara, SSC RAS Publ., 2007, 582 p.