УДК 681.5 DOI: 10.18698/2541-8009-2018-02-253

АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ И МАШИННОЕ ОБУЧЕНИЕ

М.Г. Хачатрян 5019973@mail.ru
SPIN-код: 3633-3934

П.И. Чепик p.chepik@bmstu.net
SPIN-код: 3691-4451

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Выполнен обзор литературы, посвященной анализу социальных сетей. Кратко рассмотрены основные направления анализа социальных сетей, сформулированы цели анализа и некоторые задачи в этих направлениях. Подробно описан основной метод представления социальной сети, а также приведены основные понятия, применяемые в большинстве направлений анализа социальных сетей. Рассмотрены сферы применения методов машинного обучения в анализе социальных сетей и случаи, в которых их использование целесообразно. Для демонстрации принципа работы методов машинного обучения приведен пример решения задачи идентификации спама методом машинного обучения с учителем.

Ключевые слова

Социальные сети, машинное обучение, алгоритм, структура сети, графы, модель сети, классификация полярностей, идентификация спама

Поступила в редакцию 11.01.2018

Введение. Социальной сетью называют структуру, состоящую из массива узлов, которые представлены социальными объектами и набором связей между ними. Социальными объектами могут быть люди, организации, страны, веб-страницы. Связи представляют собой взаимодействия между этими объектами. Анализ социальных сетей широко распространен в различных сферах: разведывательных и правоохранительных мероприятиях (например, обнаружение противоправной деятельности в киберпространстве [1]), банковской сфере (кредитный скорринг), предпринимательской деятельности (таргетированная реклама), политическом секторе (анализ мнений) и т. д.

Основные направления анализа социальных сетей. Существует множество направлений анализа социальных сетей, однако среди них можно выделить четыре основных [2]: структурное, ресурсное, нормативное и динамическое. При структурном анализе основное внимание уделяют форме сети и интенсивности взаимодействий, при ресурсном каждого участника сети рассматривают как источник ресурсов и анализируют возможность достижения той или иной цели с помощью данных участников. При нормативном анализе изучают социальные роли, уровни доверия между участниками и процессы их взаимодействий. При динамическом анализе исследуют изменения в сетевой структуре с течением времени, в том числе причины возникновения и исчезновения связей. Кроме того, в данных направлениях существует множество задач [3], таких

как определение сообществ в социальных сетях, статистический анализ, анализ содержания социальной сети и т. д.

Представление сетей. Поскольку социальные сети могут иметь различные структуры и размеры, единого способа их представления во всех сферах применения не существует. Однако некоторые понятия являются базовыми для многих сфер [4].

Узлы и графы. Узлы в социальной сети представляют собой множество значений $N = \{1, \ldots, n\}$, которые связаны между собой некоторыми отношениями. В различных сферах применения узел называют по-разному: агент, актор, игрок.

Граф (N, E) — это математический объект, содержащий множество узлов $N = \{1, \ldots, n\}$ и совокупность пар узлов E, называемых ребрами. При этом в классическом ненаправленном графе выполняются следующие условия:

1) каждая пара объектов из N встречается в E не более одного раза;

2) пары из E являются неупорядоченными;

3) в графе нет петель: $(x, x) \notin E$.

Модель сети. Модель сети можно определить как некоторый объект — заместитель объекта-оригинала, где в данном случае объектом-оригиналом является социальная сеть. Применение моделей вызвано тем, что сеть может иметь чрезвычайно сложную структуру и для ее анализа эффективнее использовать математическую модель этой сети, описывающую главные элементы структуры сети. С помощью правильно выбранной модели можно довольно эффективно имитировать поведение этой сети и анализировать ее поведение в течение времени [5].

Сообщества. При анализе социальных сетей наибольший интерес представляют именно связи в социальной сети, поскольку бóльшая часть метрик сети (показателей, используемых при анализе социальных сетей) будет зависеть от этих связей [6]. Для образования связей в социальных сетях огромную роль играют сообщества. Сообществом можно назвать множество людей с одинаковыми интересами, при этом люди, состоящие в сообществе, могут оказаться совершенно не связанными между собой (они могут быть из разных социальных слоев, иметь разную работу, географическое положение и т. д.). Сообщество является движущей силой для образования связей, именно поэтому в современных виртуальных социальных сетях предусмотрена возможность построения сообществ, в противном случае эти социальные сети ничем не отличались бы от телефонных справочников [7].

Машинное обучение. Основная причина, по которой необходимо использовать алгоритмы машинного обучения, — это неструктурированный характер данных в социальной сети. Конечно, некоторые данные легко сортировать, заносить в базу данных и классифицировать, однако большинство данных, таких как сообщения пользователей, являются неструктурированными и требуют более тщательного анализа [8].

Одной из популярнейших задач анализа социальных сетей является задача классификации полярностей. Она заключается в том, чтобы определить, к какому типу можно отнести то или иное высказывание (негативное, положитель-

ное, нейтральное). Для решения этой задачи на данный момент применяют три алгоритма: машинное обучение с учителем, частичное машинное обучение с учителем и без учителя. Более подробно про применение данных подходов для анализа настроений можно прочитать в статьях [9–11].

В качестве примера алгоритма машинного обучения с учителем рассмотрим алгоритм Байеса, использовавшийся для идентификации спама в [12]. В данной работе в качестве исследуемой социальной сети был выбран Twitter. Для оценки сообщения на наличие спама использовано два вида критериев: на основе связей и на основе сообщений. В критерии на основе связей рассматривалось количество людей (аккаунтов) N_{fo} на которых подписан конкретный пользователь, количество людей, которые находятся в друзьях у пользователя N_{fr} и коэффициент следования $r_{ff} = \dfrac{N_{fo}}{N_{fo} + N_{fr}}$. В качестве критерия на основе сообщений пользователя было выбрано расстояние Левенштейна для различных сообщений одного пользователя, которое определяется как минимальное количество операций вставки, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

В основе наивного классификатора Байеса лежит теорема Байеса:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)},$$

где в данном контексте $P(Y \mid X)$ — вероятность того, что аккаунт X принадлежит к классу Y (спам или не спам); $P(X \mid Y)$ — вероятность встретить аккаунт X во всех аккаунтах класса Y; $P(Y)$ — безусловная вероятность встретить аккаунт класса Y среди всех аккаунтов; $P(X)$ — безусловная вероятность встретить аккаунт X среди всех аккаунтов.

Согласно классификатору Байеса, каждый аккаунт представляет собой набор свойств аккаунта, не зависящих друг от друга. Поэтому условную вероятность встречи аккаунта можно представить в виде

$$P(Y \mid X) = \frac{\prod_{i=1}^{d} P(X \mid Y),\ P(Y)}{P(X)},$$

где d — число свойств аккаунта X.

В зависимости от того, какой из классов Y имеет максимальное значение, принимают решение о том, является ли аккаунт X источником спама.

Выводы. В настоящей статье описаны основные направления и понятия в анализе социальных сетей. Для представления социальных сетей применяют элементы теории графов, а для имитации их поведения используют различные модели, выбираемые в зависимости от структуры сети и поставленной задачи. В ходе проведенного обзора литературы по анализу социальных сетей выявлено,

что методы машинного обучения наиболее широко распространены в сферах, где необходимо работать с неструктурированным набором данных, таким как сообщения пользователей в виртуальных социальных сетях. Наиболее перспективным направлением является анализ мнений пользователей в этих сообщениях.

Литература

[1] Басараб М.А., Иванов И.П., Колесников А.В., Матвеев В.А. Обнаружение противоправной деятельности в киберпространстве на основе анализа социальных сетей: алгоритмы, методы и средства (обзор). *Вопросы кибербезопасности*, 2016, № 4(17), с. 11–19.

[2] Чураков А.Н. Анализ социальных сетей. *СоцИс*, 2001, № 1, с. 109–121.

[3] Батура Т.В. Модели и методы анализа компьютерных социальных сетей. *Программные продукты и системы*, 2013, № 3, pp. 130–137.

[4] Jackson M.O. *Social and economic networks*. Princeton University Press, 2010, 520 p.

[5] Newman M. *Networks: an introduction*. Oxford University Press, Oxford, 2010, 720 p.

[6] Wu M. Social network analysis 101. URL: https://community.lithium.com/t5/Science-of-Social-Blog/Social-Network-Analysis-101/ba-p/5706 (дата обращения 06.11.2017).

[7] Wu M. From weak ties to strong ties: community vs. social networks 3. URL: https://community.lithium.com/t5/Science-of-Social-Blog/From-Weak-Ties-to-Strong-Ties-Community-vs-Social-Networks-3/ba-p/6834 (дата обращения 06.11.2017).

[8] Huddy G. How does machine learning improve social media analysis? URL: https://www.crimsonhexagon.com/blog/machine-learning-social-media-analysis/ (дата обращения 13.11.2017).

[9] Habernal I., Ptácek T., Steinberger J. Sentiment analysis in Czech social media using super-vised machine learning. *Proc. 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2013, pp. 65–74.

[10] Huang Z., Zhao Z., Liu Q., Wang Z. An unsupervised method for short-text sentiment analysis based on analysis of massive data. *Intelligent computation in big data era*. Springer, 2015, pp. 169–176.

[11] He X., Zhang H., Chao W., Wang D. Semi-supervised learning on cross-lingual sentiment analysis with space transfer. *Proc. IEEE First Int. Conf. on Big Data Computing Service and Applications*, 2015, pp. 371–377.

[12] Wang A.H. Detecting spam bots in online social networking sites: a machine learning approach. *Proc. 24th annual IFIP WG 11.3 working conf. on Data and applications security and privacy*. Springer, 2010, pp. 335–342.

Хачатрян Микаэл Гагикович — студент кафедры «Информационная безопасность», МГТУ им. Н. Э. Баумана, Москва, Российская Федерация.

Чепик Полина Игоревна — студентка кафедры «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Ключарев Петр Георгиевич, кандидат технических наук, доцент кафедры «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

ANALYSIS OF THE SOCIAL NETWORKING SITES AND THE MACHINE LEARNING

M.G. Khachatryan 5019973@mail.ru
SPIN-code: 3633-3934

P.I. Chepik p.chepik@bmstu.net
SPIN-code: 3691-4451

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

This work presents a literature review regarding the social networking sites analysis. We briefly consider the principal directions of the social networking sites analysis and set out the objectives of this analysis and some tasks in these directions. The article thoroughly describes the basic method of the social networking website representation and also presents the basic terms applied in most directions of the social networking sites analysis. We consider the application fields for the machine learning methods in the social networking sites analysis and the cases where their application is reasonable. To demonstrate the working principle of the machine learning methods, we provide an example of solving the spam identification problem by means of the machine learning method with the teacher.

Keywords

Social networking sites, machine learning, algorithm, network structure, columns, network model, classification of polarities, spam identification

References

[1] Basarab M.A., Ivanov I.P., Kolesnikov A.V., Matveev V.A. Detection of illegal activities in cyberspace on the basis of the social networks analysis: algorithms, methods, and tools (a survey). *Voprosy kiberbezopasnosti* [Cybersecurity issues], 2016, no. 4(17), pp. 11–19.

[2] Churakov A.N. Social network analysis. *SotsIs* [Sociological Studies], 2001, no. 1, pp. 109–121.

[3] Batura T.V. Social networks analysis models and methods. *Programmnye produkty i sistemy* [Software & Systems], 2013, no. 3, pp. 130–137.

[4] Jackson M.O. *Social and economic networks*. Princeton University Press, 2010, 520 p.

[5] Newman M. Networks: an introduction. Oxford University Press, Oxford, 2010, 720 p.

[6] Wu M. Social network analysis 101. Available at: https://community.lithium.com/t5/Science-of-Social-Blog/Social-Network-Analysis-101/ba-p/5706 (accessed 06 November 2017).

[7] Wu M. From weak ties to strong ties: community vs. social networks 3. Available at: https://community.lithium.com/t5/Science-of-Social-Blog/From-Weak-Ties-to-Strong-Ties-Community-vs-Social-Networks-3/ba-p/6834 (accessed 06 November 2017).

[8] Huddy G. How does machine learning improve social media analysis? Available at: https://www.crimsonhexagon.com/blog/machine-learning-social-media-analysis/ (accessed 13 November 2017).

[9] Habernal I., Ptácek T., Steinberger J. Sentiment analysis in Czech social media using super-vised machine learning. *Proc. 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2013, pp. 65–74.

[10] Huang Z., Zhao Z., Liu Q., Wang Z. An unsupervised method for short-text sentiment analysis based on analysis of massive data. Intelligent computation in big data era. Springer, 2015, pp. 169–176.

[11] He X., Zhang H., Chao W., Wang D. Semi-supervised learning on cross-lingual sentiment analysis with space transfer. *Proc. IEEE First Int. Conf. on Big Data Computing Service and Applications*, 2015, pp. 371–377.

[12] Wang A.H. Detecting spam bots in online social networking sites: a machine learning approach. *Proc. 24th annual IFIP WG 11.3 working conf. on Data and applications security and privacy*. Springer, 2010, pp. 335–342.

Khachatryan M.G. — student, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Chepik P.I. — student, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Klyucharev P.G., Cand. Sc. (Eng.), Assoc. Professor, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.