КЛАСТЕРИЗАЦИЯ МЕДИАКОНТЕНТА ИЗ СОЦИАЛЬНЫХ СЕТЕЙ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ BIGDATA

И.А. Рыцарев[1], Д.В. Кирш[1, 2], А.В. Куприянов[1, 2]

[1] Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Россия, г. Самара, Московское шоссе, д. 34,

[2] ИСОИ РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, 443001, Россия, Самарская область, г. Самара, ул. Молодогвардейская, д. 151

Аннотация

Статья посвящена одной из ключевых проблем, возникающих при анализе социальных сетей, – проблеме классификации учётных записей на основе медиаконтента, загружаемого пользователями. Основными трудностями на пути решения проблемы являются гетерогенность контента (как по формату, так и по содержанию) и колоссальные объёмы анализируемой информации, что приводит к чрезмерной вычислительной сложности её обработки, а зачастую и к полной неэффективности традиционных методов анализа. В статье мы обсуждаем подход к кластеризации медиаконтента из социальных сетей на основе текстового аннотирования с использованием технологии BigData – современного и эффективного инструмента, позволяющего решить проблемы обработки данных большого объёма. Для проведения вычислительных экспериментов была собрана большая выборка разнородных изображений (фотографии, картины, поздравительные открытки и т. д.) из реальных профилей пользователей социальной сети Twitter. Проведённое исследование подтвердило высокое качество кластеризации медиаконтента, в среднем, значение ошибки составило порядка 5 %.

Ключевые слова: кластеризация, технология BigData, текстовое аннотирование, социальные сети, анализ медиа-контента, алгоритм k-means, GoogLeNet.

Цитирование: **Рыцарев, И.А.** Кластеризация медиаконтента из социальных сетей с использованием технологии BigData / И.А. Рыцарев, Д.В. Кирш, А.В. Куприянов // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 921-927. – DOI: 10.18287/2412-6179-2018-42-5- 921-927.

Введение

Бурный рост социальных сетей стал абсолютно закономерным процессом современности: каждый день пользователи генерируют сотни терабайт медиаконтента (в основном изображения и видео). Анализ такого рода данных имеет большое значение во многих отраслях. Например, невозможно переоценить влияние интернет-маркетинга на продвижение товаров и услуг на рынке. Но для наиболее эффективного использования данных механизмов необходимо чётко понимать запросы пользователя. Публикации, объявления, взаимодействие с различными сообществами в социальных сетях становятся основным источником информации об интересах и потребностях пользователя [1].

На сегодняшний день социальные сети по своему масштабу представляют собой ещё одну модель реальности, в которой строится взаимодействие людей. Причём явление интернет-коммуникации достигло таких масштабов, что изучать его традиционными методами социологической науки уже не представляется возможным. Устаревают традиционные методы прогнозирования развития общества и государства – неопределённость становится неизбежной частью бытия. В сложившейся ситуации разработка универсальных методов анализа медиаконтента из социальных сетей, несомненно, является актуальной задачей, решение которой имеет крайне важное социальное значение.

Методы интеллектуального анализа данных являются наиболее эффективным и широко используемым средством, позволяющим аналитикам обрабатывать терабайты цифровой информации и выявлять индивидуальные особенности целевой аудитории. Существует множество алгоритмов, которые решают конкретные задачи: обнаружение лиц [2], выбор области интереса [3], идентификация объекта [4] и многие другие. Однако проблема глобальной классификации пользователей является более общей и связана с анализом миллиардов неоднородных медиафайлов [5,6].

Одним из негативных последствий возникшего многообразия стало существенное размывание границ между классами (а в ряде случаев полное смешение), которые мы привыкли выделять при анализе медиаконтента. Вследствие чего особо остро встала проблема кластеризации. Целью настоящего исследования стала разработка технологии анализа медиаконтента на основе формирования кластеров объектов, отличающихся общностью контента. В качестве исследуемых медиаобъектов были выбраны изображения, размещённые пользователями в социальных сетях в открытом доступе.

Предложенная в данной статье технология представляет полный цикл от сбора медиаконтента из социальных сетей и до проведения его анализа с использованием технологии BigData.

Сбор данных из социальных сетей

Задачу выявления необходимой информаций из социальных сетей зачастую разделяют на следующие этапы: сбор, предобработка, анализ данных и визуализация результатов (рис. 1). При этом наибольшей информативностью обладают данные, генерируемые в режиме «онлайн», что приводит к ряду серьёзных проблем с хранением такого рода данных [7].

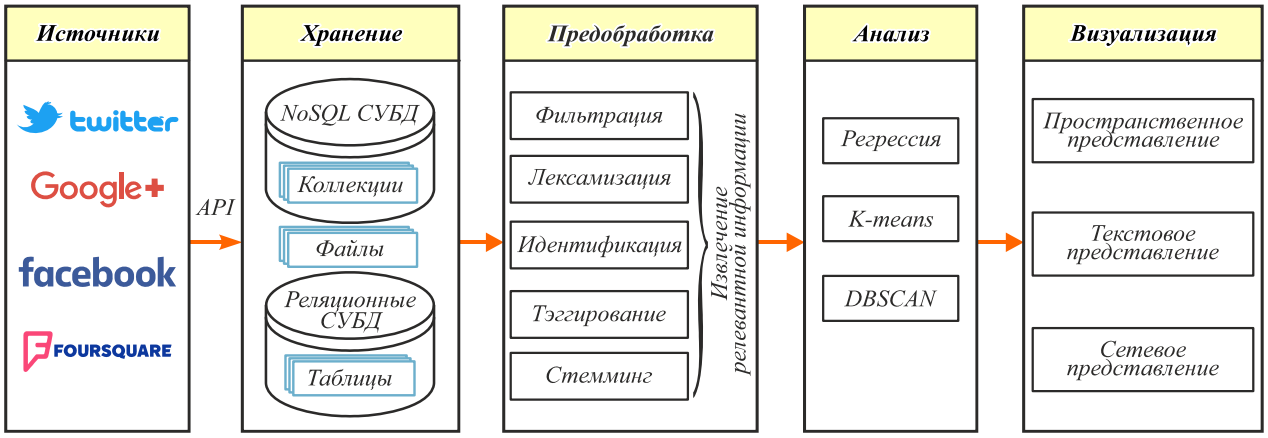


Рис. 1. Общая схема предложенного алгоритма кластеризации изображений

С другой стороны, фильтрация медиаконтента на начальном этапе играет решающую роль в уменьшении объёма данных для последующего анализа. Такой отбор может производиться на основе анализа метаданных: например, по определённой геолокации или по типу опубликованного контента [8].

В настоящее время всё большее число социальных сетей (Twitter, Google+, Facebook, Foursquare и др.) предоставляет открытый программный интерфейс (API), который не только позволяет встраивать функциональность социальной сети в другие приложения, но и автоматизирует процедуру сбора пользовательских данных, опубликованных в открытом доступе.

В данной работе в качестве источника медиаконтента была выбрана социальная сеть Twitter. Данный выбор обусловлен следующими причинами:

1. Сеть обеспечивает открытый доступ к данным, хранящимся на серверах, без строгих ограничений на скачивание.
2. Она представляет из себя вторую по популярности социальную сеть после Facebook среди пользователей по всему миру. Однако последняя не обеспечивает открытый доступ к своим данным.
3. Twitter не является проблемно-ориентированной сетью и, следовательно, отражает общественное мнение максимально широкого круга пользователей [9].

Сбор данных из социальной сети Twitter может осуществляться с помощью программ Apache Ambari и Flume, данный метод был подробно описан в нашей предыдущей работе [10]. Однако для сбора данных с использованием ряда фильтров зачастую более удобно разрабатывать специализированный программный продукт с использованием стандартных библиотек программирования (twitter4j, tweepy и др.) [11].

На рис. 2 показана схема алгоритма сбора данных из социальной сети. Она включает в себя два принципиально разных фильтра (F1 и F2). Первый фильтр (F1) направлен на извлечение и анализ метаданных (тип, геолокация, тэги и т.д.). Второй фильтр (F2) представляет собой крайне сложную процедуру аннотирования. Иными словами, преобразование медиаконтента в набор текстовой информации, которая описывает его содержимое.

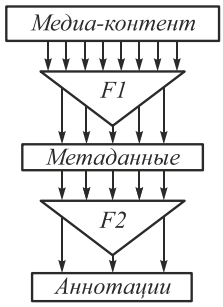


Рис. 2. Обобщённая схема алгоритма сбора данных социальной сети

Кластеризация данных изображения с использованием технологии BigData

Кластеризация изображений с использованием традиционных подходов требует огромных вычислительных ресурсов и занимает много времени [12]. В связи с этим предлагается использовать преимущества технологии BigData и специализированного программно-аппаратного комплекса обработки данных, принадлежащего Самарскому университету, позволяющие многократно ускорить анализ данных большого объёма по сравнению с традиционными настольными системами [13]. В состав комплекса входят:

1. Программно-аппаратный комплекс для хранения и аналитического анализа структурированных данных – IBM Puredata for Analytics (Netezza) с объёмом дискового пространства не менее 96 ТБ (с сжатием 4 раза и полной репликацией данных).
2. Hadoop-кластер для распределенного хранения и аналитической обработки неструктурированных данных – сервер управления IBM x3630 M4 (2 процессора Intel Xeon E5-2450v2; 96 ГБ оперативной памяти, 2 жёстких диска объёмом 600 ГБ) и четыре сервера обработки данных IBM x3630 M4 (2 процессора Intel Xeon E5 2450v2; 96 ГБ оперативной памяти, объём доступного хранилища – 8 ТБ).

В качестве инструмента предобработки медиаконтента посредством аннотирования (фильтр F2) было решено использовать нейронную сеть GoogLeNet, результатом работы которой является вектор вероят-

ностей принадлежности изображения каждому из 1000 классов (определённых в результате работы исследовательской группы) [14]. На рис. 3 представлен пример изображения из социальной сети, относящегося к классу «Postcards» (рис. 3*а*), а также рассчитанный по нему вектор вероятностей (рис. 3*б*) (для большей наглядности классы с вероятностями менее 0,03 были исключены). Из диаграммы видно, что анализируемая фотография, скорее всего, относится к классу 567 и 804 – два наиболее ярко выраженных пика.

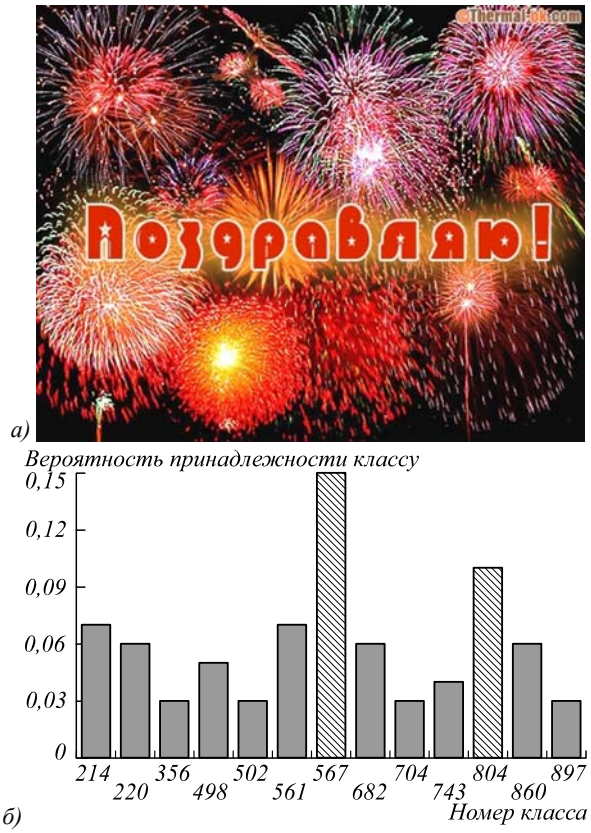


Рис. 3. Пример анализа изображения класса Postcards

На рис. 4 показан аналогичный пример работы алгоритма для изображения из социальной сети, относящегося к классу «Animals». Представленная диаграмма позволяет сделать вывод, что анализируемое изображение относится к классам 698 и 840.

Следующим этапом работы являлась кластеризация полученных векторов. В качестве метода кластеризации был применён метод k-means++ [15]. Данный алгоритм обеспечивает значительно лучшее качество классификации, чем стандартный алгоритм k-means, за счёт оптимизации выбора начальных условий.

Программная реализация кластерной части выполнялась на языке программирования высокого уровня Python с использованием программной платформы Spark для распределённой обработки данных. Вычислительные эксперименты проводились на высокопроизводительном кластере для обработки данных сверхбольшого объёма.

Исследования проводились на большом наборе медиа-контента (около 230 000 изображений с метаданными), открыто опубликованного пользователями социальной сети Twitter и собранного с использованием предоставленного API.

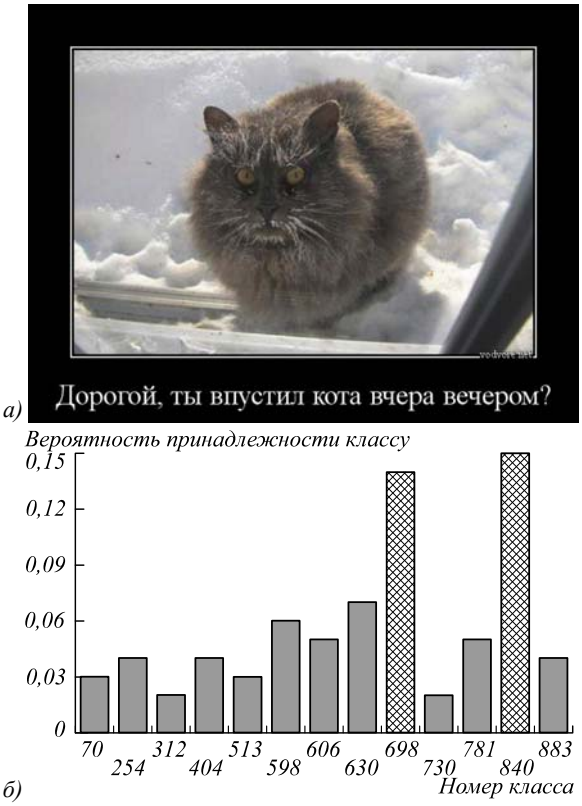


Рис. 4. Пример анализа изображения класса Animals

Результатом проведённых экспериментов стало определение 13 основных классов изображений, наиболее часто публикуемых пользователями социальной сети Twitter, а также распределение всех изображений по данным кластерам. Список основных классов изображений с их кратким описанием представлен в табл. 1.

Для вышеупомянутых классов изображений мы вычислили частоту распределения всех изображений по 13 выделенным кластерам (табл. 2).

На основании представленных данных можно выделить три ведущих класса изображений, которые чаще всего публикуются в социальной сети Twitter: фотографии, персоналии и открытки. Эти классы изображений зачастую не являются тематическими и наиболее характерны для социальных сетей (VK, Facebook и др.). Таким образом, можно предположить, что большинство представленных изображений не являются оригинальными и представляют собой пересылку данных пользователей из разных социальных сетей.

Качество кластеризации изображений

Следующим этапом исследований стала оценка качества предлагаемого метода кластеризации изображений в социальных сетях с использованием большого набора собранных данных. Однако, как можно видеть из табл. 2 распределение всех изображений по классам является неравномерным. Для устранения отмеченного недостатка, было выбрано случайным образом по 5000 изображений каждого класса (65 000 изображений всего).

Табл. 1. Характеристики основных классов изображений

Номер	Класс	Описание
1	Photo	Групповые фотографии, портреты
2	Animals	Представители животного мира
3	Sport	Спортивные соревнования, спортивный инвентарь
4	Auto / Moto	Автомобили, мотоциклы и другие транспортные средства
5	Selfie	Фотографии людей на фронтальную камеру мобильного телефона
6	Text	Изображения с большим количеством текста
7	Plants	Растения крупным планом
8	Water	Изображения, на которых присутствует вода
9	Postcards	Поздравительные открытки
10	Monochromatic	Изображения с монохромным фоном
11	Equipment	Технические устройства
12	Building	Архитектурные сооружения
13	Other	Изображения, не попавшие ни в один из предыдущих классов

Табл. 2. Распределение изображений по 13 основным классам

Имя класса	Количество изображений, шт.
Photo	21 163
Animals	15 330
Sport	9 924
Auto / Moto	18 379
Selfie	31 981
Text	23 959
Plants	13 466
Water	17 877
Postcards	27 825
Monochromatic	9 975
Equipment	11 482
Building	12 097
Other	13 236

Сравнение предложенной технологии проводилось с другим существующим подходом к анализу медиаконтента, опубликованного в работе [16] и основанного на оценке гистограмм яркостей. Результаты эксперимента представлены в табл. 3.

Табл. 3. Значение ошибки кластеризации изображения для 13 основных классов

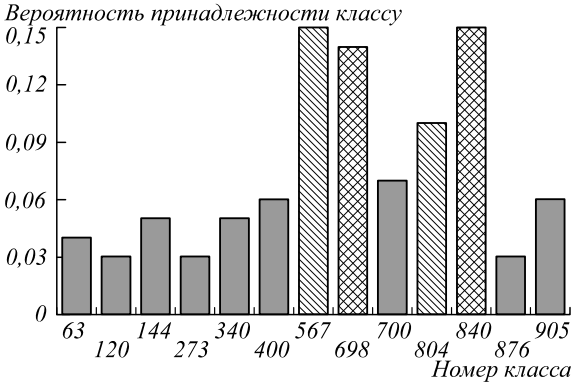
Номер класса	Ошибка алгоритма на основе сети GoogLeNet, %	Ошибка алгоритма на основе оценки гистограмм яркостей, %
Photo	5	7
Animals	3	5
Sport	7	3
Auto / Moto	4	8
Selfie	2	12
Text	1	1
Plants	2	4
Water	3	1
Postcards	10	18
Monochromatic	12	20
Equipment	8	11
Building	6	9
Other	5	5

В среднем, значение ошибки кластеризации алгоритмом на основе сети GoogLeNet по всем классам составило чуть более 5 %, а максимальное значение – порядка 12 %, в то же время, значения ошибки кластеризации алгоритмом на основе оценки гистограмм яркостей составили 8 % и 20 % соответственно.

В заключение, мы провели исследование состоятельности оценки, получаемой с использованием предложенной технологии. Для этого среди оценок вероятностей принадлежности классам, построенных для всего множества изображений, была найдена та, характер которой в наибольшей степени повторяет особенности оценок вероятности двух изображений, представленных ранее. Не сложно заметить, что на распределении вероятностей (рис. 5*а*) присутствуют как два пика, встречавшихся на примере из класса «Postcards» (567, 804), так и два пика, соответствующие примеру из класса «Animals» (698, 840). Следовательно, если технология работает корректно, само изображение также должно обладать одновременно чертами двух заинтересовавших нас классов. Смелая гипотеза, но так ли это на самом деле? Достаточно взглянуть на изображение (рис. 5*б*), по которому было построено распределение, чтобы мгновенно убедится в правильности сделанного предположения.

Вывод

Мы представили технологию кластеризации контента в социальных сетях на основе алгоритмов аннотации классов GoogLeNet и k-means++. Предложенный метод показал многообещающие результаты и позволил соотнести каждое изображение с одним из 13 классов, описывающих наиболее часто размещаемый медиаконтент. Серия экспериментов доказала высокое качество кластеризации, в среднем, ошибка составила порядка 5 %, для сравнения, алгоритм на основе оценки гистограмм продемонстрировал ошибку порядка 8 %. Данный факт показывает, что использование свёрточных нейронных сетей позволяет значительно повысить точность классификации и кластеризации изображений. Дальнейшие исследования будут направлены на более подробный анализ медиа-контента, а также на более широкий охват существующих социальных сетей.



б)

Рис. 5. Пример анализа изображения, имеющего общие черты классов Postcards и Animals

Благодарности

Работа выполнена при частичной поддержке Министерства науки и высшего образования РФ в рамках выполнения работ по Государственному заданию ФНИЦ «Кристаллография и фотоника» РАН (соглашение № 007-ГЗ/Ч3363/26); Министерства образования и науки РФ в рамках реализации мероприятий Программы повышения конкурентоспособности Самарского университета среди ведущих мировых научно-образовательных центров на 2013–2020 годы; грантов РФФИ № 15-29-03823, № 16-41-630761, № 17-01-00972, № 18-37-00418; в рамках госзадания по теме № 0026-2018-0102 «Оптоинформационные технологии получения и обработки гиперспектральных данных».

Литература

1. **Maxwell, D.** Crisees: Real-time monitoring of social media streams to support crisis management / D. Maxwell, S. Raue, L. Azzopardi, C.W. Johnson, S. Oates. – In: Advances in information retrieval / ed. by R. Baeza-Yates, A.P. de Vries, H. Zaragoza, B.B. Cambazoglu, V. Murdock, R. Lempel, F. Silvestri. – Berlin: Springer, 2012. – P. 573-575. – DOI: 10.1007/978-3-642-28997-2_68.
2. **Scott, J.** Social network analysis / J. Scott. – 3rd ed. – London: Sage Publications Ltd, 2017. – 216 p. – ISBN: 978-1-4462-0904-2.
3. **Borgatti, S.P.** Analyzing social networks / S.P. Borgatti, M.G. Everett, J.C. Johnson. – 2nd ed. – London: Sage Publications Ltd, 2013. – 384 p. – ISBN: 978-1-5264-0410-7.
4. **Kirsh, D.V.** 3D crystal structure identification using fuzzy neural networks / D.V. Kirsh, O.P. Soldatova, A.V. Kupriyanov, I.A. Lyozin, I.V. Lyozina // Optical Memory & Neural Networks (Information Optics). – 2017. – Vol. 26, Issue 4. – P. 249-256. – DOI: 10.3103/S1060992X17040026.
5. **Marra, F.** Blind PRNU-based image clustering for source identification / F. Marra, G. Poggi, C. Sansone, L. Verdoliva // IEEE Transactions on Information Forensics and Security. – 2017. – Vol. 12, Issue 9. – P. 2197-2211. – DOI: 10.1109/TIFS.2017.2701335.
6. **Xu, X.** SCAN: a structural clustering algorithm for networks / X. Xu, N. Yuruk, Z. Feng, T.A.J. Schweiger // Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. – 2007. – P. 824-833. – DOI: 10.1145/1281192.1281280.
7. **Khotilin, M.I.** Visualization and cluster analysis of social networks / M.I. Khotilin, A.V. Blagov // CEUR Workshop Proceedings. – 2016. – Vol. 1638. – P. 843-850. – DOI: 10.18287/1613-0073-2016-1638-843-850.
8. **Semertzidis, K.** How people describe themselves on Twitter / K. Semertzidis, E. Pitoura, P. Tsaparas // Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks. – 2013. – P. 25-30. – DOI: 10.1145/2484702.2484708.
9. **Blagov, A.** Big data instruments for social media analysis / A. Blagov, I. Rytsarev, K. Strelkov, M. Khotilin // Proceedings of the 5th International Workshop on Computer Science and Engineering. – 2015. – P. 179-184.
10. **Rytsarev, I.** Creating the model of the activity of social network Twitter users / I. Rytsarev, A. Blagov // Journal of Telecommunication, Electronic and Computer Engineering (JTEC). – 2017. – Vol. 9, Issues 1-3. – P. 27-30.
11. **Rytsarev, I.A.** Development and research of algorithms for clustering data of super-large volume / I.A. Rytsarev, A.V. Blagov // CEUR Workshop Proceedings. – 2017. – Vol. 1903. – P. 80-83.
12. **Dhanachandra, N.** Image segmentation using K-means clustering algorithm and subtractive clustering algorithm / N. Dhanachandra, K. Manglem, Y.J. Chanu // Procedia Computer Science. – 2015. – Vol. 54. – P. 764-771. – DOI: 10.1016/j.procs.2015.06.090.
13. **Kazanskiy, N.** Performance analysis of real-time face detection system based on stream data mining frameworks / N. Kazanskiy, V. Protsenko, P. Serafimovich // Procedia Engineering. – 2017. – Vol. 201. – P. 806-816. – DOI: 10.1016/j.proeng.2017.09.602.
14. **Szegedy, C.** Going deeper with convolutions / C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. – 2015. – 9 p. – DOI: 10.1109/CVPR.2015.7298594.
15. **Bahmani, B.** Scalable k-means++ / B. Bahmani, B. Moseley, A. Vattani, R. Kumar, S. Vassilvitskii // Proceedings of the VLDB Endowment. – 2012. – Vol. 5, Issue 7. – P. 622-633. – DOI: 10.14778/2180912.2180915.
16. **Rejito, J.** Image indexing using color histogram and k-means clustering for optimization CBIR in image database / J. Rejito, A.S. Abdullahi, Akmal, D. Setiana, B.N. Ruchjana // Journal of Physics: Conference Series. – 2017. – Vol. 893, Issue 1. – 012055. – DOI: 10.1088/1742-6596/893/1/012055.

Сведения об авторах

Рыцарев Игорь Андреевич, 1993 года рождения, в 2017 окончил магистратуру Самарского университета по специальности «Прикладная математика и информатика», работает инженером на кафедре технической кибернетики. Область научных интересов: обработка данных социальных сетей, большие данные. E-mail: *rycarev_igoryan@mail.ru* .

Кирш Дмитрий Викторович, 1990 года рождения. В 2018 году окончил аспирантуру Самарского университета по направлению «Информатика и вычислительная техника». В настоящее время ассистент кафедры технической кибернетики Самарского национального исследовательского университета имени академика С.П. Королева; стажёр-исследователь лаборатории математических методов обработки изображений Института систем обработки изображений РАН – филиала ФНИЦ «Кристаллография и фотоника» РАН. Сфера научных интересов: цифровая обработка изображений и распознавание образов; методы описания и сравнения кристаллических решёток; классификация кристаллических решёток. E-mail: *limitk@mail.ru* .

Куприянов Александр Викторович, профессор кафедры технической кибернетики Самарского национального исследовательского университета имени академика С.П. Королёва; старший научный сотрудник лаборатории математических методов обработки изображений Института систем обработки изображений РАН. Сфера научных интересов: цифровая обработка сигналов и изображений; распознавание образов и искусственный интеллект; анализ и интерпретация биомедицинских сигналов и изображений. E-mail: *alexkupr@gmail.com* .

ГРНТИ: 28.23.15.
Поступила в редакцию 24 октября 2018 г. Окончательный вариант – 30 октября 2018 г.

CLUSTERING OF MEDIA CONTENT FROM SOCIAL NETWORKS USING BIGDATA TECHNOLOGY

I.A. Rytsarev [1], D.V. Kirsh [1,2], A.V. Kupriyanov [1,2]
[1] Samara National Research University, Moskovskoye Shosse 34, 443086, Russia, Samara,
[2] IPSI RAS – Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardeyskaya 151, 443001, Russia, Samara

Abstract

The article deals with one of the key problems of the social network analysis – the problem of classifying accounts based on media content uploaded by users. The main difficulties are the content heterogeneity (both in format and subject) and the large volumes of data, which leads to excessive computational complexity of its processing and often to the complete inefficiency of traditional analysis methods. In the article, we discuss an approach to the clustering of media content from social networks based on textual annotation using BigData technology – a modern and efficient tool that allows to solve the problem of large data volume processing. To carry out computational experiments, a large sample of heterogeneous images (photographs, paintings, postcards, etc.) was collected from real Twitter accounts. The results confirmed the high quality of media content clustering, the average error was around 5 %.

Keywords: cluster analysis, BigData technology, text annotation, social networks, media content analysis, k-means clustering, GoogLeNet.

Citation: Rycarev IA, Kirsh DV, Kupriyanov AV. Clustering of media content from social networks using BigData technology. Computer Optics 2018; 42(5): 921-927. DOI: 10.18287/2412-6179-2018-42-5-921-927.

Acknowledgements: This work was partially supported by Ministry of Science and Higher Education within the State assignment FSRC "Crystallography and Photonics" RAS (Agreement No 007-ГЗ/Ч3363/26); by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of Samara University among the world's leading scientific and educational centers for 2013-2020 years; by the Russian Foundation for Basic Research grants (# 15-29-03823, # 16-41-630761, # 17-01-00972, # 18-37-00418); in the framework of the state task # 0026-2018-0102 "Optoinformation technologies for obtaining and processing hyperspectral data".

References

[1] Maxwell D, Raue S, Azzopardi L, Johnson CW, Oates S. Crisees: Real-time monitoring of social media streams to support crisis management. In Book: Baeza-Yates R, de Vries AP, Zaragoza H, Cambazoglu BB, Murdock V, Lempel R, Silvestri F, eds. Advances in information retrieval. Berlin: Springer; 2012: 573-575. DOI: 10.1007/978-3-642-28997-2_68.

[2] Scott J. Social network analysis. 3rd ed. London: Sage Publications Ltd; 2017. ISBN: 978-1-4462-0904-2.

[3] Borgatti SP, Everett MG, Johnson JC. Analyzing social networks. 2nd ed. London: Sage Publications Ltd; 2018. ISBN: 978-1-5264-0410-7.

[4] Kirsh DV, Soldatova OP, Kupriyanov AV, Lyozin IA, Lyozina IV. 3D crystal structure identification using fuzzy neural networks. Opt Mem Neural Networks 2017; 26(4): 249-256. DOI: 10.3103/S1060992X17040026.
[5] Marra F, Poggi G, Sansone C, Verdoliva L. Blind PRNU-based image clustering for source identification. IEEE Transactions on Information Forensics and Security 2017; 12(9): 2197-2211. DOI: 10.1109/TIFS.2017.2701335.
[6] Xu X, Yuruk N, Feng Z, Schweiger TAJ. SCAN: a structural clustering algorithm for networks. Proc 13th ACM SIGKDD international conference on Knowledge discovery and data mining 2007: 824-833. DOI: 10.1145/1281192.1281280.
[7] Khotilin MI, Blagov AV. Visualization and cluster analysis of social networks. CEUR Workshop Proceedings 2016; 1638: 843-850. DOI: 10.18287/1613-0073-2016-1638-843-850.
[8] Semertzidis K, Pitoura E, Tsaparas P. How people describe themselves on Twitter. Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks. ACM 2013: 25-30. DOI: 10.1145/2484702.2484708.
[9] Blagov A, Rytsarev I, Khotilin M, Strelkov K. Big data instruments for social media analysis. Proceedings of the 5th International Workshop on Computer Science and Engineering 2015: 179-184.
[10] Rytsarev I, Blagov A. Creating the model of the activity of social network Twitter users. Journal of Telecommunication, Electronic and Computer Engineering 2017; 9(1-3): 27-30.
[11] Rytsarev IA, Blagov AV. Development and research of algorithms for clustering data of super-large volume. CEUR Workshop Proceedings 2017; 1903: 80-83.
[12] Dhanachandra N., Manglem K., Chanu Y. J. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science 2017; 54: 764-771. DOI: 10.1016/j.procs.2015.06.090.
[13] Kazanskiy N, Protsenko V, Serafimovich P. Performance analysis of real-time face detection system based on stream data mining frameworks. Procedia Engineering 2017; 201: 806-816. DOI: 10.1016/j.proeng.2017.09.602.
[14] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. Proc IEEE conference on Computer Vision and Pattern Rrecognition 2015. DOI: 10.1109/CVPR.2015.7298594.
[15] Bahmani B, Moseley B, Vattani A, Kumar R, Vassilvitskii S. Scalable k-means++. Proc VLDB Endowment 2012; 5(7): 622-633. DOI: 10.14778/2180912.2180915.
[16] Rejito J, Abdullahi AS, Akmal, Setiana D, Ruchjana BN. Image indexing using color histogram and k-means clustering for optimization CBIR in image database. Journal of Physics: Conference Series 2017; 893(1): 012055. DOI: 10.1088/1742-6596/893/1/012055.

Authors' information

Igor Andreevich Rytsarev (b. 1993) is a postgraduate student of Samara University. He graduated (2017) with a master's degree in Applied Mathematics and Informatics. At present, he is an engineer of Technical Cybernetics department. The area of interests includes social network analysis and Big Data technology. E-mail: *rycarev_igoryan@mail.ru* .

Dmitriy Victorovich Kirsh (b. 1990) completed (2018) the postgraduate program in Computer Science and Computer Engineering. At present, he is a lecturer at the Technical Cybernetics department of Samara University and also a junior researcher at the IPSI RAS – Branch of the FSRC "Crystallography and Photonics" RAS. The area of interests includes digital image processing and pattern recognition, methods of mathematical formulation and comparison of crystal lattices, classification of crystal lattices. E-mail: *limitk@mail.ru* .

Alexandr Victorovich Kupriyanov (b. 1978) graduated (2001) from the S.P. Korolyov Samara State Aerospace University (SSAU). He received his PhD in Technical Sciences (2004). At present, he is a senior researcher at the Image Processing Systems Institute of the Russian Academy of Sciences, and holding a part-time position of Associate Professor at Technical Cybernetics department of Samara University. The area of interests includes digital signals and image processing, pattern recognition and artificial intelligence, biomedical imaging and analysis. His list of publications contains more than 80 scientific papers, including 35 articles and 1 monograph published. E-mail: *alexkupr@gmail.com* .

Received October 24, 2018. The final version – October 30, 2018.