Кластеризация изображений социальных сетей с использованием технологии BigData

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

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация. Настоящая работа посвящена одной из ключевых проблем, возникающих в процессе анализа данных пользователей социальных сетей, – проблеме классификации пользователей на основе загружаемых изображений. Основными затруднениями при решении данной задачи являются: разнородный характер изображений (фотографии, художественные работы, поздравительные открытки и т.д.) и колоссальные объёмы анализируемой информации, что приводит к чрезмерной вычислительной сложности её обработки. В настоящей работе рассматривается подход к кластеризации изображений на основе классовой аннотации, использующий технологию BigData – современное и эффективное средство для борьбы с указанными затруднениями. В качестве исследуемых данных для проведения вычислительных экспериментов использовалась обширная выборка изображений, собранных с реальных профилей социальной сети Twitter.

1. Введение

В настоящее время социальные сети переживают бурный рост: каждый день их пользователи генерируют сотни терабайт медиа – изображений и видео. Их анализ имеет огромное значения во многих сферах бизнеса. К примеру, невозможно переоценить влияние интернет-маркетинга на продвижение товаров и услуг на рынке. Однако, для эффективного использования данных механизмов необходимо чётко понимать запросы пользователей. Источником такой информации как раз и могут служить материалы, публикуемые пользователями социальных сетей, а также формируемые в результате их обмена связи между пользователями и целые сообщества. Таким образом, рассматриваемая в рамках данной работы задача кластеризации изображений, публикуемых пользователями социальной сети Twitter, с использованием классовой аннотации нейронной сети GoogleNet является, несомненно, актуальной задачей, решение которой имеет также большое научное значение в сфере анализа данных.

2. Сбор данных с социальных сетей

Источником данных для исследования была выбрана социальная сеть Twitter. Это было сделано по следующим причинам:

- сеть предоставляет открытый доступ к своим данным (нет ограничения на доступ к данным сервера);
- является второй по популярности социальной сетью (после Facebook, который не предоставляет открытый доступ к своим данным) среди пользователей во всем мире;
- Twitter не является предметной сетью, а значит, отражает общественное мнение более широкого круга пользователей [4].

Сбор данных социальной сети Twitter может осуществляться посредством программных продуктов Apache Ambari и Flume, подробнее данный метод описан в работе[5]. Однако для сбора данных с применением ряда фильтров, зачастую, удобнее разработать свой программный продукт с использованием стандартных библиотек (twitter4j, tweepy и т.п.) [6].

В рамках данного исследования был разработан программный комплекс на языке программирования Python, содержащий модуль авторизации, модуль сбора данных и модуль фильтрации. Данный программный комплекс позволяет собирать данные по геолокации, по ключевым словам, по пользователю, а также кэшировать все медиафайлы пользователя. Для исключения перебоев в работе программного комплекса, связанных с превышением лимитов установленных социальной сетью Twitter, в него внедрено множество ключей авторизации. Программный комплекс работает в режиме real-time мониторинга, а также может делать запросы на получение расположенной на серверах информации.

За время работы программного комплекса было собрано более 120 000 изображений с аккаунтов пользователей.

3. Кластеризация изображений с использованием технологии BigData.

Кластеризация изображений с использованием традиционных подходов требует огромных вычислительных ресурсов и занимает длительное время. В связи с этим, было принято решение воспользоваться технологией BigData и вычислительным кластером для обработки данных сверхбольшого объема, имеющимся в наличии у Самарского Университета. В качестве первого этапа кластеризации изображений было решено использовать текстовую аннотацию изображений. Для этого был взят результат алгоритма текстовой аннотации изображений посредством использования нейронной сети GoogleNet, которая на выходе дает вектор вероятностей принадлежности изображения к каждому из 1000 классов, определённых в результате работы исследовательской группы [7]. Пример результата работы алгоритма представлен на рисунке 1.

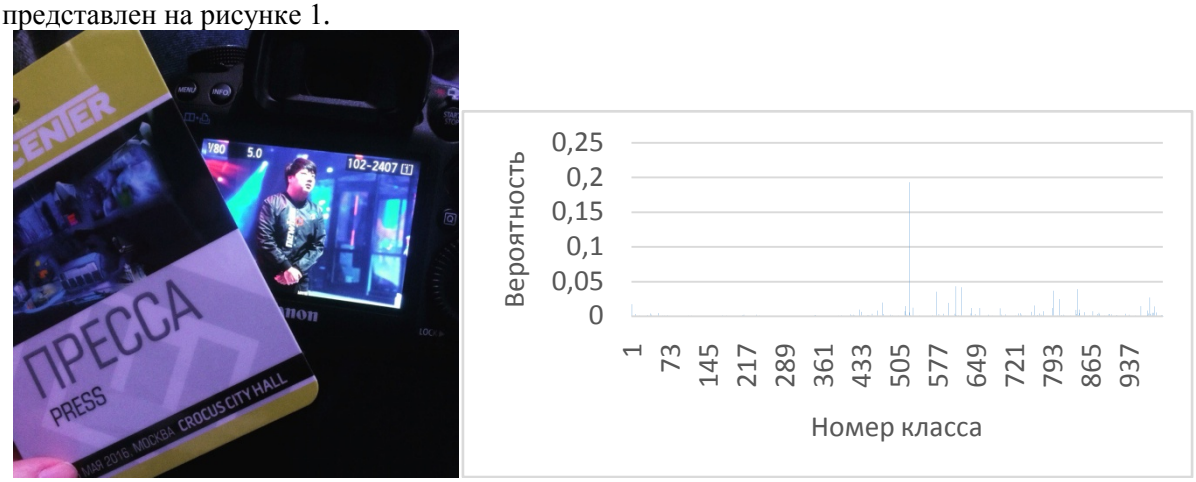

Рисунок 1. Пример загруженного изображения и гистограммы его принадлежности к рассматриваемым классам изображений.

Следующим этапом работы стала кластеризация полученных векторов. В качестве метода кластеризации был взят метод k-means с косинусным расстоянием в качестве меры схожести. Программная реализация кластерной части была выполнена на языке программирования высокого уровня Python с использованием программной платформы распределённой обработки

данных Spark. Проведение вычислительных экспериментов осуществлялось на высокопроизводительном кластере лаборатории по обработке данных сверхбольшого объема Самарского Университета.

Результатом проведённых экспериментов стало выявление 13 основных классов изображений, наиболее часто публикуемых пользователями социальной сети Twitter, и распределение всех изображений по данным кластерам.

Таблица 1. Характеристика основных классов изображений

Номер	Класс	Описание
1.	Фотографии	Фотографии групп: семьи, людей, рисунки людей
2.	Животные	Фотографии животных, рисунки животных
3.	Спорт	Фотографии со спортивных мероприятий, спортинвентарь, фотографии с природы
4.	Авто/мото	Фотографии автомобилей, мотоциклов и других транспортных средств
5.	Селфи	Фотографии на фронтальную камеру
6.	Текст	Картинки с крупным текстом
7.	Растения	Картинки с растениями крупным планом
8.	Вода	Фотографии/картинка на которых присутствует вода
9.	Открытки	Поздравления с праздниками
10.	Фотографии с однотонным фоном	Изображения с однотонным задним планом
11.	Техника	Технические устройства
12.	Здания	Фотографии зданий
13.	Прочее	Прочие картинки/изображения

Для перечисленных классов изображений была построена гистограмма, наглядно демонстрирующая характер распределения всех изображений по выделенным 13 кластерам (рисунок 2).

На представленной диаграмме особо выделяются три лидирующих класса изображений, наиболее часто публикуемых в социальной сети Twitter: фотографии, селфи и открытки. Данные классы изображений, зачастую, не являются тематическими и наиболее часто встречаются в полноформатных социальных сетях (ВКонтакте, Facebook и т.д.). В связи с этим, можно сделать предположение о том, что большинство представленных изображений не являются оригинальными и представляют собой репосты пользователей из полноформатных социальных сетей.

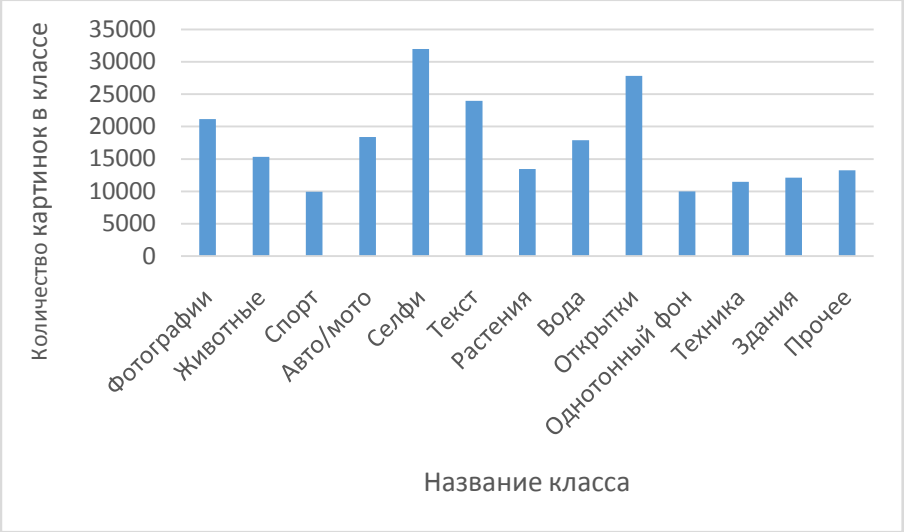

Рисунок 2. Гистограмма распределения изображений по основным кластерам.

4. **Заключение**

В работы были апробированы алгоритмы GoogleNet и K-means для кластеризации изображений социальной сети Twitter на основе классовой аннотации. В результате было получено 13 классов изображений. В дальнейшем результаты работы могут быть использованы в разработке алгоритмов кластеризации изображений с использованием технологии Big Data.

5. Благодарности

Работа выполнена при поддержке Федерального агентства научных организаций (соглашение № 007-ГЗ/Ч3363/26); Министерства образования и науки РФ в рамках реализации мероприятий Программы повышения конкурентоспособности СГАУ среди ведущих мировых научно-образовательных центров на 2013–2020 годы; грантов РФФИ № 15-29-03823, № 15-29-07077, № 16-41-630761, № 16-29-11698, № 17-01-00972, № 18-37-00418; программы № 6 фундаментальных исследований ОНИТ РАН «Биоинформатика, современные информационные технологии и математические методы в медицине» 2017 г.

6. Литература

[1] 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.

[2] 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.

[3] Xu, X. Scan: a structural clustering algorithm for networks / X. Xu et al. // Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2007. – P. 824-833.

[4] Blagov, A. Big Data Instruments for Social Media Analysis / A. Blagov, I. Rytcarev, K. Strelkov, M. Khotilin // Proceedings of the 5th International Workshop on Computer Science and Engineering. – 2015. – P. 179-184.

[5] 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(1-3). – P. 27-30.

[6] 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.

[7] Szegedy, C. Going deeper with convolutions / C. Szegedy et al. // Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. – P. 1-9.

Clustering of images in social media with the use of BigData technology

I.A. Rytsarev[1], A.V. Kupriyanov[1,2], D.V. Kirsh[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. This work is devoted to one of the key problems arising in the analysis of social media – the problem of account classification on the basis of images uploaded by users. The main difficulties in solving the problem are the heterogeneous nature of images (photos, artworks, greeting cards, etc.) and colossal volumes of analyzed information, which leads to excessive computational complexity of its processing. In the paper, we discuss an approach to image clustering based on class annotation, using BigData technology – a modern and effective tool to handle the described difficulties. To carry out computational experiments, a large sample of images from real profiles of Twitter users was collected.

Keywords: Social networks, Twitter, Image clustering, GoogleNet.