Clustering of social media content with the use of BigData technology

I A Rytsarev1, A V Kupriyanov1,2, D V Kirsh1,2 and K S Liseckiy1

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

e-mail: rycarev@gmail.com

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 media uploaded by users. The main difficulties in solving the problem are the heterogeneous nature of the content (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 social media clustering based on class annotation, using BigData technology – a modern and effective tool to handle the described difficulties. To carry out computational experiments, we collected a large sample of images from real profiles of Twitter users.

1. Introduction

Today, online social networks are experiencing rapid growth: every day their users generate hundreds of terabytes of media content (mostly images and videos). The analysis of this data is of great importance in many areas of business. For example, it is impossible to overestimate the impact of Internet marketing on the promotion of goods and services in the market. For effective use of these mechanisms, it is necessary to clearly understand the user’s requests. Publications, announcements, interaction with different communities in social networks become the major source of such information [1].

Methods of data mining can help marketers analyze terabytes of social images to understand the audience in a more personal way. There are a lot of algorithms that solves particular tasks: face detection [2,3], region of interest selection [4,5], object identification [6,7], etc. However, the problem of account classification is more general and connected with the analysis of billions heterogeneous images [8]. In this paper we propose a technique for clustering of images in social media, including a data collection procedure and the use of BigData technology.
2. Data collection from social networks

The task of gathering the necessary information from social networks can be divided into data collection, filtering, and processing. The greatest value for obtaining information is represented by data generated on-line [9]. At the same time, filtering can take place either by a certain geolocation of interest, or by the subject matter of the published content [10].

The social network Twitter was chosen as the source of the data for the study. This was done for the following reasons:

- The network provides open access to its data (there is no restriction on access to server data);
- It is the second most popular social network (after Facebook, which does not provide open access to its data) among users around the world;
- Twitter is not a problem-specific network, and therefore reflects the public opinion of a wide range of users [11].

Data collection of Twitter social network can be carried out through Apache Ambari and Flume software, this method is described in more detail in our previous work [12]. However, to collect data using a number of filters, it is often more convenient to develop a specialized software product using standard programming libraries (twitter4j, tweepy, etc.) [13].

The algorithm of data collection developed in the framework of the research, in addition to the designated filtering, uses an extra element of user feedback by means of an automatic request for clarifying the information about published messages. Figure 1 shows the scheme of the algorithm for collecting the data from the social network.

![Figure 1. Generalized scheme of the algorithm for collecting social network data.](image)

On-line data generated by the social network is passed through two filters. The F1 filter analyzes an image metadata, in particular, the specified geolocation. The second filter F2 is configured for keywords or topics and analyzes the image itself.

In the framework of this study, a program complex was developed in the Python programming language containing an authorization module, a data collection module, and a filtering module. This software allows to collect data by geolocation, by keywords, by user, in addition, it caches all user’s media files. To avoid interruptions in the work of the program complex, which are associated with exceeding the limits set by the social network Twitter, many authorization keys have been embedded into it. The software complex operates in real-time monitoring mode, and can also make requests for information located on the servers.

The received data contain a huge amount of information: message text, metadata, links to attachments (videos, photos), information about message views, number of reposts, user information, and others. This allows to conduct a lot of research analyzing user profiles, activity, message texts, etc. The data are represented in the form of JSON-object (Table 1). An example of the collected data is shown in Figure 2. For experimental study, more than 120,000 images were collected from user accounts.
Table 1. The structure of JSON-response.

| Field         | Description                                           |
|---------------|-------------------------------------------------------|
| 'id'          | message id                                           |
| 'text'        | message test                                         |
| 'user'.id'    | author id                                            |
| 'user'.name'  | author name                                           |
| 'user'.location' | author location                                    |
| 'user'.geo_enabled' | geolocation status (enabled/disabled)     |
| 'user'.lang'  | author language                                      |
| 'place'.id'   | location id                                           |
| 'place'.name' | location name                                         |
| 'place'.bounding_box' | location coordinates                      |
| 'coordinates' |                                                       |

Figure 2. An example of the data collected from the social network.

3. Clustering of image data using BigData technology

Clustering of image data using traditional approaches requires huge computing resources and takes a long time. In this regard, we decided to use the advantages of BigData technology and the specialized software-hardware data processing complex for processing large amounts of information, owned by Samara University. The complex consists of:

- software and hardware complex for storage and analytical analysis of structured data – IBM Puredata for Analytics (Netezza) with the volume disk space of at least 96TB (with 4x compression and full replication of data);
- Hadoop-cluster for distributed storage and analytical processing unstructured data – IBM management server x3630 M4 (2x Intel Xeon Processor E5-2450v2; 96 GB of RAM; 2x 600GB HDD memory) and four IBM data processing servers x3630 M4 (2x Intel Xeon Processor E5 2450v2; 96 GB of RAM; 8TB HDD memory).

As the first stage, it was decided to use the text annotation of images. We chose the neural network GoogleNet, which on the output gives the vector of probabilities of the image belonging to each of the 1000 classes (determined as a result of the work of the research group) [14]. An example of the result of the algorithm is presented in Figure 3. It shows the analyzed photo and the filtered probability.
distribution diagram (classes with probabilities <0.01 are not displayed). From the diagram it is clear that the photo under analysis most likely belongs to the class 520.

Figure 3. Example 1: loaded image and its class distribution.

Figure 4 shows another example of the result of the algorithm. It shows that the analyzed photo belongs to the classes 512, 536, 630 and 750.

Figure 4. Example 2: loaded image and its class distribution.

The next stage of the work is the clustering of the resulting vectors. As a method of clustering, the k-means method was taken. The algorithm of k-means clusterization consists of the following steps:

1. Choose $K$ initial centroids.
2. Determine the distance of each data item to the centroid (cosine distance):

\[ d_{ik} = \frac{\sum_{j} x_{ik} c_{jk}}{\sqrt{\sum_{i} x_{ik}^2} \sqrt{\sum_{j} c_{jk}^2}}. \]

3. Form $K$ clusters by assigning the data items to the closest centroid.
4. Re-compute the centroid of each cluster.
5. Repeat step 3 and step 4 until the centroids do not change their positions.

The aim of the k-means algorithm is to minimize the squared error function:

\[ J = \sum_{i=1}^{K} \sum_{j=1}^{N} d_{ij}^2. \]

The software implementation of the cluster part was performed in a high-level Python programming language using the Spark software platform for distributed data processing. Computational experiments were carried out on the high-performance cluster for processing data of a very large volume.
The results of the experiments were the identification of 13 main classes of images, most frequently published by users of the social network Twitter, and the distribution of all images on these clusters. A list of main classes of images with a brief description is presented in Table 2.

Table 2. Characteristics of the main classes of images.

| Number | Class    | Description                                              |
|--------|----------|----------------------------------------------------------|
| 1      | Photo    | Photos of groups: families, people, drawings of people   |
| 2      | Animals  | Photos of animals, drawings of animals                   |
| 3      | Sport    | Photos from sporting events, sports equipment, photos from nature |
| 4      | Auto / Moto | Pictures of cars, motorcycles and other vehicles      |
| 5      | Selfie   | Photos on the front camera                               |
| 6      | Text     | Pictures with large text                                 |
| 7      | Plants   | Pictures with plants close-up                            |
| 8      | Water    | Photos / pictures on which water is present              |
| 9      | Postcards| Greetings on holidays                                    |
| 10     | Monochromatic | Photos with monochrome background             |
| 11     | Equipment| Technical devices                                        |
| 12     | Building | Photos of buildings                                      |
| 13     | Other    | Other pictures / images                                  |

For the above mentioned image classes, we calculated a histogram that visually demonstrates the distribution of all images across the selected 13 clusters (Figure 5).

Figure 5. Distribution of social media over 13 main classes.

The presented diagram highlights three leading classes of images, most often published on the social network Twitter: photos, selfies and postcards. These classes of images are often not thematic and are most frequently found in social networks (VK, Facebook, etc.). In this regard, we can make the assumption that most of the images presented are not original and represent reposts of users from different social networks.

4. Quality of the image clusterization

To evaluate the quality of the proposed technique for clusterization of images in social media, we conducted a series of computational experiments using a big set of collected data (120 000 images with metadata). At each iteration after training the algorithm, 1000 images from the set were randomly selected for clusterization. The results of the experiment are shown in Figure 6. The average error of the clusterization did not exceed 6 %, whereas the overall error was less than 12 %.
5. Conclusion

We presented a technique for clustering social media content based on GoogleNet and K-means algorithms of class annotation. The technique has shown promising results and allowed to assign each image to one of the 13 classes describing most frequently posted media content. The series of experiments proved the high quality of clusterization, the average error was less than 6%. Further research will be aimed at more detailed analysis of media content from social networks with the use of the developed algorithms.

6. References

[1] Maxwell D, Raue S, Azzopardi L, Johnson C and Oates S 2012 Crises: Real-time monitoring of social media streams to support crisis management Advances in Information Retrieval (Springer Berlin Heidelberg) 573-575
[2] Sergeev A E, Konushin A S, Konushin V S 2016 Reducing background false positives for face detection in surveillance feeds Computer Optics 40(6) 958-967 DOI: 10.18287/2412-6179-2016-40-6-958-967
[3] Kalinovskii I A, Spitsyn V G 2016 Review and testing of frontal face detectors Computer Optics 40(1) 99-111 DOI: 10.18287/2412-6179-2016-40-1-99-111
[4] Ilyasova N, Paringer R, Kupriyanov A, Kirsh D 2017 Intelligent feature selection technique for segmentation of fundus images The Seventh International Conference on Innovative Computing Technology (INTECH 2017) 129-134
[5] Ilyasova N, Shirokanev A, Kirsh D, Paringer R, Kupriyanov A, Zamycky E 2017 Development of coagulate map formation algorithms to carry out treatment by laser coagulation Procedia Engineering 201 271-279
[6] Soldatova O P, Lyozin I A, Lyozina I V, Kupriyanov A V, Kirsh D V 2015 Application of fuzzy neural networks for defining crystal lattice types in nanoscale images Computer Optics 39(5) 787-794
[7] Kupriyanov A V, Kirsh D V 2015 Estimation of the Crystal Lattice Similarity Measure by Three-Dimensional Coordinates of Lattice Nodes Optical Memory & Neural Networks (Information Optics) 24(2) 145-151
[8] Xu X 2007 Scan: a structural clustering algorithm for networks Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM) 824-833
[9] Khotilin M I and Blagov A V 2016 Visualization and Cluster Analysis of Social Networks CEUR Workshop Proceedings 1638 843-850
[10] Semertzidis K, Pitoura E and Tsaparas P 2013 How people describe themselves on Twitter Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks 25-30
[11] Blagov A, Rytcarev I, Strelkov K and Khotilin M 2015 Big Data Instruments for Social Media Analysis Proceedings of the 5th International Workshop on Computer Science and Engineering 179-184
[12] Rytsarev I and Blagov A 2017 Creating the Model of the Activity of Social Network Twitter Users *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 9(13) 27-30

[13] Rytsarev I A and Blagov A V 2017 Development and research of algorithms for clustering data of super-large volume *CEUR Workshop Proceedings* 1903 80-83

[14] Szegedy C 2015 Going deeper with convolutions *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1-9

**Acknowledgement**

This work was partially supported 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".