Creating the Model of the Activity of Social Network Twitter Users

Igor Rytsarev, Aleksandr Blagov
Samara State Aerospace University, Samara, 443086, Russia.
rycarev@gmail.com

***Abstract*—The present article is dedicated to the research in the area of analyzing text data from social network Twitter. Due to large volumes of data generated in social networks, the collection and processing of these data can be performed by means of methods and instruments of Big Data. The article describes the process of determining the most popular words and terms in social network Twitter. Based on the results drawn from the analysis of their usage, a model of user's activity has been developed. In addition, mathematical statistics methods were used to validate the adequacy of the model.**

***Index Terms*—Twitter; Big Data; Statistical Model; Social Network.**

I. Introduction

In information technologies, high-volume data sets mean data sets that have the size above the abilities of typical databases for adding, storing, managing and analyzing information [1]. There are many series of approaches, instruments and methods of processing structured and unstructured high-volume data. These series include tools of mass-parallel processing of vaguely structured data, such as NoSQL, MapReduce algorithms, the Hadoop project frameworks and libraries [2, 3]. Precisely, these instruments were chosen for data collection and analysis.

At present, social networks are at the peak of popularity: Now millions of people are using Facebook and Twitter. The direction of Big Data, which is related to social media is one of the most perspective areas, and it is developing dynamically. According to [4-7], it could also shape the attitude of researches in this area.

Many companies need to analyze data collected from social networks, for example to assess the relationship of users to their products [8]. Besides, the analysis of these data can also be used to solve security questions [9]. The results of analyzing data from social networks can be used to determine the ratio of users to a particular topic and the relationship between users.

The present article describes the research carried out for determining the most popular words and terms in social network Twitter and building a conditional model of users' activity.

II. Methods and Materials

The process of working with high-volume data consists of three main stages:

i. Data collection,
ii. Data processing,
iii. Analysis of the processed data.

The creation of models or prognosis can take place during the process of solving a number of tasks based on the results of the analysis. This can also be allocated as the fourth stage.

Nowadays, many leading companies in the world, such as IBM, Oracle, Microsoft and others, are trying to solve the problem of working with high-volume data and offering their own solutions for the data collection and processing [10]. The most popular instruments for collecting and processing data are: Apache Hadoop, Apache Ambari, Biginsights, Cloudera, Hortonworks, Storm [11-13]. In this article, the Apache Ambari tool was used for data collection. This tool was installed and configured on the cluster of the laboratory of "High-volume data sets processing" of SSAU. It is able to collect continuous parallel data stream during a large period of time and in large quantities.

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

i. Twitter is the second most popular network among users from the whole world. However, in contrast to Facebook, which is the first most popular social network, Twitter offers free access to the networks' data. Hence, there are no restrictions on the access to the streams of data from the server.
ii. The users of this social network are basically exchanging information, particularly text information, which is an advantage during processing.
iii. Twitter is not a specialized network, and it most widely reflects public opinion on many issues of interest.

The Apache Ambari has its own tool, called Flume, and its main function is for collecting text data (logs, messages, service information) [14]. This tool was used to collect necessary information. Thus, collecting text data from social network Twitter can be produced in real-time mode based on specific parameters. However, stream data, collected from social networks, contain a lot of service information. For the future analysis, only those data sets, which are interesting are considered important, so, it is necessary to separate the service information from the information, which is needed.

The structured data was derived by arranging and excluding the service information, which are not necessary, using the MapReduce technology [15]. After that, using a developed program tool, the specific information, which is needed for the further analysis and creating the mathematical model, was gathered. The whole stage of unstructured data processing is shown in the Figure 1.

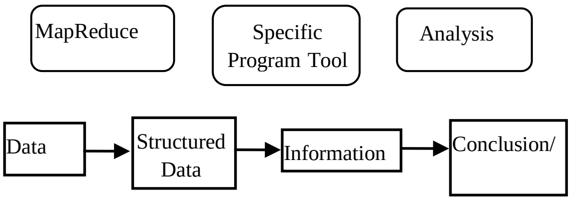

Figure 1: The scheme of unstructured data processing

The process of cleaning the data using MapReduce working consists of two main steps: Map step and Reduce step.

The pre-processing of the input data takes place during the Map step. For this, one of the computers (which is called the “master node”) receives the input data, partitions them into parts and transmits them to other computers (called the “worker node”). This step was named by the same-name of higher-order function.

At the Reduce-step, the convolution of pre-processing data takes place. The master node receives answers from worker nodes and based on them, it forms the result – the solution of the problem, which was originally formulated.

During the research, the cluster was deployed, and Hortonworks Sandbox tool was configured. After that, the SQL-request cut the whole “system” information and left only the “helpful” fields, in the form of written data.

In order to extract from the necessary information from the structured data, a specific program tool (based on high-level language Java) was developed. This tool “chose” all the “necessary” fields from the structured data, such as: language, text, time zone, and the time of creation.

The total count of tweets K_L for each location L (country) is:

$$K_L = \sum_i (k_i \in L), \tag{1}$$

where k_i – each next tweet from the processing stream. The frequency of use $Count(w)$ of each unique word w is determined from the whole sets S of text data:

$$Count(w) = \sum_i (k_i \in S). \tag{2}$$

The sentiment of each tweet is determined from the dictionary - d, in which the sentiment was described:

$$sp(w,d) = \begin{cases} 0, & if \;\; w \, is \; negative, \\ 1, & if \;\; w \, is \; neutral, \\ 2, & if \;\; w \, is \; positive. \end{cases} \tag{3}$$

The analysis of structured data based on the set of structured data from social networks contains a lot of important and useful information: location, time, sentiment, etc. Using various mathematical methods, including statistical models, many interesting conclusions can be derived, various hypotheses can be proved or disproved, and many mathematical models can be built, depending on the objectives of the work.

III. Experiments and Results

For the experiment, it was necessary to collect data set for processing all messages during the period. That is why another tool, called Flume, was configured to collect absolutely each tweet. In order to achieve this, the most frequently used articles, particles, prepositions, numerals and punctuation marks were listed in the configuration file in the field Keywords. The next step was the setting of storage (HDFS) and launching data collection. The collecting was held from 30th March 2015 to 5th April 2015 in the laboratory of “High-volume data sets processing” of Samara State Aerospace University, Samara, Russian Federation. More than 400 GB of unstructured data were collected as a result of conducting the clustering work in a week.

After the processing of the data, the following information was received. The most popular words were: «people», «love», «time», «life» and «twitter». The result was compared and analyzed for compliance with the results of the research [16], performed by researchers from Greece in 2014. The cloud of tags, which was created as a result of the working with the software system can be seen in Figure 2.

Based on each of these five words, the data collection process was conducted using Flume. These data (as well as the first time) were processed through Hortonworks Sandbox. After that, the count of tweets, which contains those words, was determined through a specific program tool

The received information was imported to Excel, and the temporal distribution of tweets (by every hour from 0.00 to 23.59), which contained the most popular words, was derived.

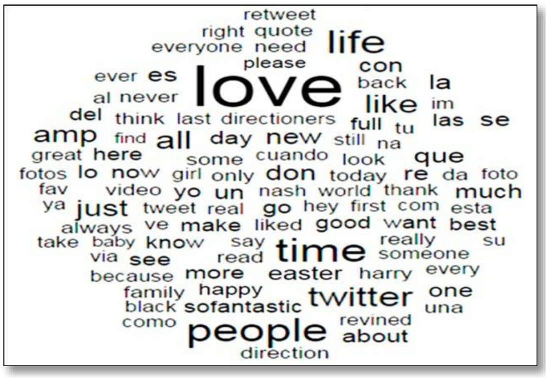

Figure 2: The cloud of the most frequently used words

For the analysis of the relative consumption of the messages within a particular cluster by locations, we will produce a normalization of each sequence by the formula:

$$x_{tn} = \frac{X_t,}{\sum_{k=0}^{23} X_k} \tag{4}$$

The schedule of distribution of this sequence can be seen in Figure 3. In order to create a model of users activity in Twitter, it is necessary to determine the sequence, which can be taken as a basis. Figure 3 shows that the sequences “People”, “Love”, “Time” and “Life” are approximately the same. Thus, the mean value of the four sequences was taken as a basis.

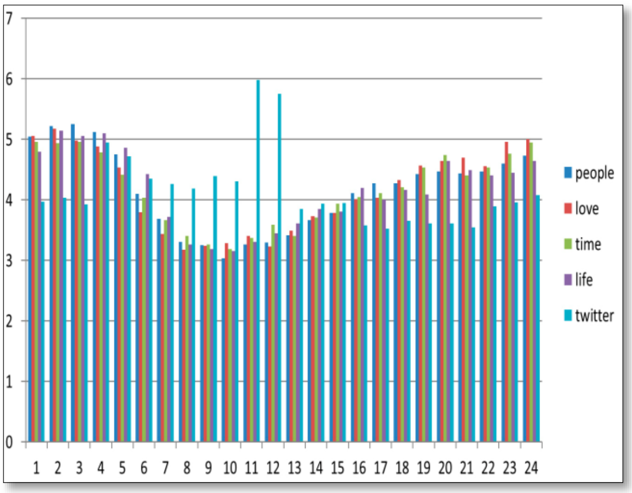

Figure 3: The normalized temporal distribution of the consumption of the tweets, which contain the most popular words

The next step is the approximation of the sequence. After that, based on the received analytic function, it is necessary to create a mathematical model, which describes the activities of Twitter's users.

The approximation was performed by a polynomial function with a power basis, using the Gram matrix and Gauss method [17].

Thereby, the received function, which approximates a temporal sequence (4) has the following form:

$$y=-0{,}000001x^6+0{,}000068x^5 -0{,}002758x^4+0{,}052427x^3- 0{,}455075x^2+1{,}379706x+3{,}917158 \quad (5)$$

Wherein, the model of the activity of users of social networks can be written as:

$$\begin{cases} X(t)=0, & t\in(-\infty;0) \\ X(t)=y, & t\in[0;24) \\ X(t)=0, & t\in[24;\infty) \end{cases}, \quad (6)$$

where y is defined by the Equation (5). The approximating curve of the model of Twitter users activity is shown in Figure 4.

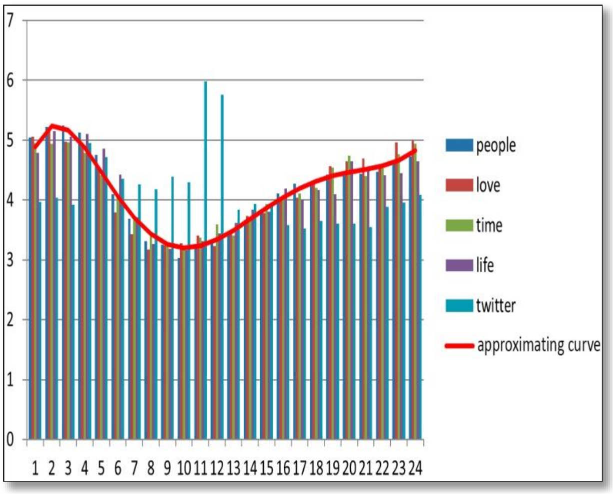

Figure 4:The approximating curve of the model of Twitter users activity

In order to test the adequacy of the model, the values of the Pearson correlation coefficient (Table 1) of the function (5) with each of the temporal sequences (4) were determined. In addition, it has been subjected to testing hypotheses of conformity of the analytical distribution of the resulting model to sequences (4) using the Kolmogorov criterion [18]. The results are shown in Table 2.

Table 1
The value of the Pearson coefficient of correlation between sequences and the model

Sequence with a keyword	The value of the Pearson correlation coefficient
people	0.979411367
love	0.966261150
time	0.982908869
life	0.937562815
twitter	0.388004818

Table 2
The Value of Coefficients of Correlation between Sequences and the Model by the Criterion of Consent of Kolmogorov

Sequence with a keyword	Λ of the criterion	Level of significance (by Kolmogorov table)
people	0.406016867	0,996
love	0.846007356	0,4806
time	0.480040795	0,9753
life	1.334817592	0,0582
twitter	7.428028772	0

IV. Discussion

The result, which was received during the research, of the most popular word, terms and notions among the users of Twitter (Figure 2) is because the large sample size may be a sufficient objective. The most popular words in this case, are also included in the list of the most popular words, defined by researchers in 2014 [16]. However, it is worth noting, that the period, during which the process of data collecting is held, can include its own characteristics in the data set. For example, in this research, one of the most popular words was also word "Easter" (the collection was a week before the holiday). It is necessary to take into account such features.

The model of the activity was created, based on data streams, which contained the most popular words: people, love, time, life, twitter. The hypothesis that the model describes the sequence, which were formed by words: people, love, time, life – can be applied with a high level of significance (Table 1 and 2). The sequence, formed by the stream of data by the word "twitter" - is not subjected to the description of the received model (6) due to its own specification.

The received results show us that the generation of the mass messages around the world has its own temporal dependence. It should be taken into account the geographical location of users for further analysis. For example, the diagram of distribution of use of tweets on the subject of «people» around the world can be seen in Figure 5.

It can be seen from this picture that the largest number of messages generated by users from the United States. The same situation by other keywords. Thus, tweets from United States have an overwhelming influence on the activity model. It can be concluded that for more specific and socially relevant models of user activity, it should be considered the locating affiliation of data streams.

The information about user activity on time has a certain social and marketing importance. Firstly, the most effective time of important and relevant information spreading is time during the peak activity of users of social networks. Secondly, this is the most suitable time for researching of the user's reaction. In this case, it is also necessary to consider the location of data streams.

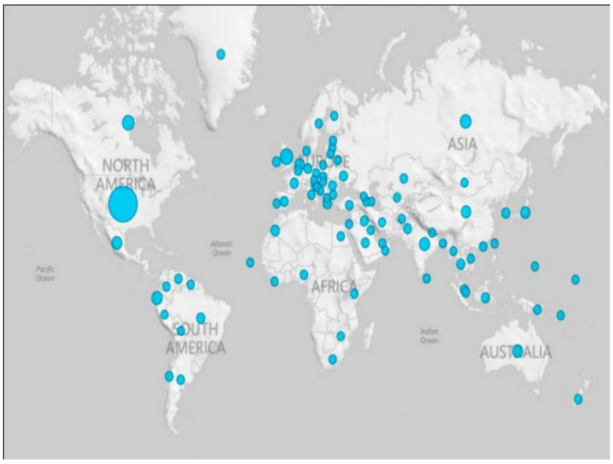

Figure 5: Diagram of distribution of use of tweets on the subject of «people» around the world

V. Conclusion

Due to their huge popularity in the whole world, social networks contain large amounts of data. One can get a slice of public opinion on various topics and issues of concern by extracting information from them. Collection and processing of this information is possible to implement using methods and tools of Big Data. Firstly, continuous work with the data in real time can be carried out, and secondly, the data can be clustered by geographical locations. Further analysis can be conducted by using methods of mathematical statistics, within these locations. Based on the results of the analysis of the information, different models, which describe these or other processes, can be created. For example: the model of activity of users of social network Twitter.

Acknowledgment

This work was 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 SSAU among the world’s leading scientific and educational centers for 2013-2020 years.

References

[1] J., Dean, S., Ghemawat. “MapReduce: simplified data processing on large clusters”, *Communications of the ACM*. T. 51. №. 1. Pp. 107-113. (2008).

[2] T.,White, *Hadoop: The definitive guide*. – " O'Reilly Media, Inc.", (2012).

[3] J.S.,Ward, A., Barker . “*Undefined by data: a survey of big data definitions*”. (2013).

[4] H., Wang, D., Can, A., Kazemzadeh, F., Bar, and S., Narayanan, “ A system for real-time twitter sentiment analysis of 2012 us presidential election cycle”. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 115-120, (2012).

[5] A., Blagov ,I., Rytcarev, K., Strelkov, M., Khotilin , “Big Data Instruments for Social Media Analysis”. *Proceedings of the 5th International Workshop on Computer Science and Engineering*, Pp. 179-184. (2015).

[6] H., Saif , Y., He, H., Alani. “Alleviating data sparsity for twitter sentiment analysis”. *CEUR Workshop Proceedings (CEUR-WS. org)*, (2012).

[7] R., Groot . “Data mining for tweet sentiment classification. (2012).

[8] W., Tan, M., Blake, M. B., Saleh, I., and S., Dustdar, “Social-network-sourced big data analytics”. *IEEE Internet Computing*. №. 5. Pp. 62-69. (2013).

[9] A., Vasilkov. How Big Data help to improve security [Electronic resource] *Computerra: 2014*. URL: http://www.computerra.ru/108760/security-n-big-data/ (accessed: 24.04.2015).

[10] A., Naydich . “Big Data: problem, technology, market” [Electronic resource] *Computerra.* (2012). URL: http://compress.ru/article.aspx?id=22725, (accessed: 14.05.2015).

[11] D., Borthakur, J., Gray, J.S., Sarma, K., Muthukkaruppan, N., Spiegelberg, H., Kuang, K., Ranganathan, D., Molkov, A., Menon, S., Rash, and R., Schmidt, “Apache Hadoop goes realtime at Facebook”. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* pp. 1071-1080, (2011).

[12] P.T., Goetz , B., O'Neill. Storm blueprints: Patterns for distributed real-time computation. – Packt Publishing Ltd, 2014.

[13] S., Wanderman-Milne ,N., Li . “Runtime Code Generation in Cloudera Impala”, *IEEE Data Eng. Bull*. T. 37. №. 1. C. 31-37, (2014).

[14] Apache Flume [Electronic resource] Hortonworks Inc. URL: http://hortonworks.com/hadoop/flume/ (accessed: 11.05.2015).

[15] J., Dean, S., Ghemawat . “MapReduce: simplified data processing on large clusters”, *Communications of the ACM*. T. 51. №. 1. Pp. 107-113.(2008).

[16] K., Semertzidis, E., Pitoura ,P., Tsaparas. “How people describe themselves on Twitter”, *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks*. ACM,.Pp. 25-30. (2013).

[17] A.A., Samarskiy, A.V., Gulin. Numerical Methods Moscow.: Science, (1989).

[18] Kolmogorov-Smirnov test [Electronic resource] // Academician. Mathematical encyclopedia [Official website]. URL: http://dic.academic.ru/dic.nsf/enc_mathematics/2279/ (accessed: 08.05.2015).