Comparative Analysis of Text Data Streams in Social Networks

Maximilian Khotilin, Aleksandr Blagov
Samara State Aerospace University, Samara, 443086, Russia
Turbomax.1994@gmail.com

***Abstract*—This article presents a research in the area of processing and analysis of text data streams in social networks. This work was conducted by means of methods and instruments of Big Data. The article details the process of working with data from social media, which starts with data collection, followed by an analysis, which subsequent ends with conclusions and hypotheses. Based on the examples of statements of the most popular sciences among users of social network Twitter, the distributions of these references taken from all over the world were developed, investigated and compared.**

***Index Terms*— Big Data; Internet; Twitter, Real-time Data; Social Networks.**

I. Introduction

Over the past decade, social networks began to play a huge role: Being the subject of the socialization of people on one hand, and being the most powerful and accessible political, ideological and economical instrument on the other hand. Due to large volumes and continuous regeneration, the research of the data from social networks can be produced by means of methods and instruments of Big Data [1-3]. The term «big data», in information technologies, means a series of approaches, instruments and methods of processing structured and unstructured high volume datasets.

In terms of marketing, social networks have become the most attractive medium for different programs' realization: They are the second place of the quickest means. In Twitter, the social network, the specialists of marketing can communicate with their audience without Service of Public Relations. Due to this, there could be a communication with a specific person, in contrast to depersonalized companies.

Using their branded terms and hash-tags, marketing specialists can learn the customers' opinion about their products, brands and companies. The global attention, which Twitter uses, shows high capabilities of social net-work technologies for public discussion and forming the perception of brands. Besides, the information collected from social networks is important in terms of questions pertaining to national security.

The direction of Big Data related to social media, is one of the most perspective and dynamically developing. The works [4], [5], [6] could also shape the attitude of the researches in this area.

Through gathering and structuring text data from social networks, the attitude of users to any selected issue can be analyzed. Additionally, through analysis, the distribution of data by countries can be received. It helps to estimate the popularity of the selected theme in specific geographical locations.

II. Methods and Materials

The algorithm of working with data sets from social networks is defined by the following scheme (Figure 1):

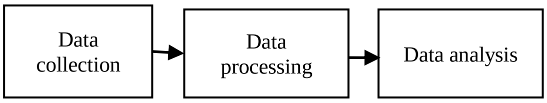

Figure 1: General diagram of working with large volumes of data

A. Data Collection

Nowadays, there are a lot of tools and programs for collecting and processing data from social networks, such as: Apache Hadoop, Apache Ambari, Biginsights, Cloudera, Hortonworks, Storm [7-9]. Each of them has its own characteristics. 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 allowed 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. But, in contrast to Facebook, which is the first most popular social network, Twitter offers free access to the networks' data. There are no restrictions on the access to the streams of data from the server.
ii. The users of this social network are exchanging, basically, only text in-formation, 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.

Thus, the collecting of text data from social network Twitter can be produced in real-time according to the set of parameters. However, the received data is unstructured.

B. The Processing of Unstructured Text Data from Social Networks

Stream data, collected from social networks, contain a lot of service information. For future analysis, considering that only those interesting data sets are considered important, it is necessary to separate the service information from the required information.

The structuring of the data was made through MapReduce technology [10], which facilitated the arrangement and exclusion of service information (the unnecessary information) from the data. After that, the specific information needed for the further analysis was gathered through a developed program tool. The whole stage of processing the unstructured data is shown in Figure 2.

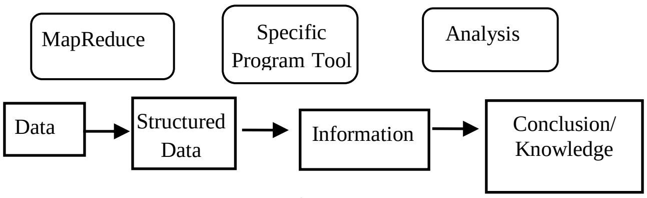

Figure 2: The scheme of unstructured data processing

MapReduse is a framework for computing some sets of distributed tasks using a large number of computers (called "nodes"), which form clusters [10].MapReduce works based on two main steps: Map step and Reduce step.

On the Map-step, the pre-processing of the input data takes place. For this, one of the computers (which is called "master node") receives the input data, and then partitions them into parts, After that it transmits them to other computers (called "worker node") for pre-processing. This step was named by the same-name higher-order function.

On the Reduce-step, the convolution of pre-processing data takes place. Master node receives answers from the 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, which cut the whole "system" information and only "helpful" fields were written.

To extract the necessary information from the structured data, a specific program tool (based on high-level language Java) was developed. This tool "chose" all "necessary" fields from the structured data, such as language, text, time zone, 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 is 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 the text data:

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

The sentiment of each tweet *sp(w,d)* is determined from the dictionary - d, in which the sentiment was described:

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

C. The Analysis of Structured Data

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

III. Experiments and Results

For the experiment, it was necessary to collect data set for processing the messages, which mentioned the words such as science, researches and discoveries. That is why another tool, called Flume, was configured to collect necessary tweets. To achieve this, words, such as «science», «Science», «SCIENCE» and many others were listed in the configuration file in the field Keywords. The next step was the setting of storage (HDFS) and launching data collection. The period of time, during which the data were collected, is seven days (a week).

The collection was held from 2nd to 7th April 2015 in the laboratory of "High volume data sets processing" of Samara State Aerospace University, Samara, Russian Federation.

The processing of the unstructured data took about a day. After the data processing, there was 920 Mb of structured information.

All of the parts of speech and hidden characters were excluding from this file. The following result was received – and the most used words were as listed below:

i. NASA;
ii. Space;
iii. Rocket;
iv. Asteroid;
v. Computer;
vi. Data;
vii. Information;
viii. Math;
ix. Medicine; and
x. Physics.

This collection of words can be divided into three clusters: «space», «computer science» and «fundamental sciences» as shown in Figure 3.

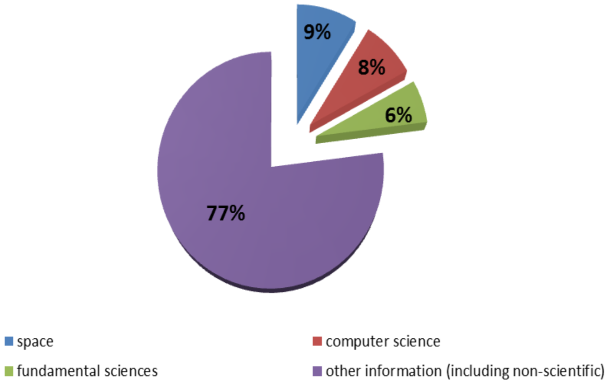

Figure 3: Diagram of the distribution of the data collected by major clusters

By the cluster «space – Sp, K_{Sp} is the count of tweets in whole text sets S:

$$K_{Sp} = \frac{\sum_i (S_i \in Sp),}{S} \tag{4}$$

By the cluster computer science – Cs, K_{Cs} is the count of tweets in whole text sets S:

$$K_{Cs} = \frac{\sum_i (S_i \in Cs),}{S} \qquad (5)$$

By the cluster fundamental science (math, medicine, physics) – F, K_F is the count of tweets in whole text sets S:

$$K_F = \frac{\sum_i (S_i \in F),}{S} \qquad (6)$$

The result suggests that the clusters, which are defined above, are the most popular science areas among the users of social networks from the whole world. Data were collected based on each of the clusters in order to establish dependencies of distributions of the collected data.

Similar procedure for collecting and processing the «science» data was also used for the collection and data processing for other clusters. This data (as well as the first time) was processed through Horton works Sandbox. After that, the count of tweets that contains those words was determined through a specific program tool.

From the received information, distributions of the use of tweets was derived which contains words, in the category «science», in cluster «space», in cluster «computer science» and in cluster «fundamental sciences», by countries. Due to differences between countries, there were three groups allocated:

i. Europe, USA, Canada and Australia – the first group;
ii. Russian Federation and Asia - the second group;
iii. South America and Africa – the third group.

An analysis of conformity of the sequences of each cluster in every group with the cluster «science» allows us to evaluate the relationship of those or other scientific areas with the category of science.

We denote X as the sequence, which corresponds to the cluster «science», Y_s as the sequence, which corresponds to the cluster «space», Y_c as the sequence, which corresponds to the cluster «computer science» and Y_f as the sequence, which correspond to the cluster «fundamental sciences».

For the analysis of the relative consumption of the messages within a particular cluster by locations, we produced a normalization of each sequence. We obtained:

$$x_n = \frac{X,}{\sum_{k=1}^{L} X_k} \qquad (7)$$

where: x_n is the normalized value of the item of the sequence science;
X is the current not normalized value of the item of the sequence science;
L is the number of elements of the sequence «science» in a particular geolocation group.

Similarly, for the sequences Y_s, Y_c, Y_f :

$$y_{S_n} = \frac{Y_{S_i},}{\sum_{k=1}^{L} Y_{S_k}}, \quad i = \overline{1, L} \qquad (8)$$

$$y_{C_n} = \frac{Y_{C_i},}{\sum_{k=1}^{L} Y_{Cs_k}}, \quad i = \overline{1, L} \qquad (9)$$

$$y_{f_n} = \frac{Y_{f_i},}{\sum_{k=1}^{L} Y_{fs_k}}, \quad i = \overline{1, L} \qquad (10)$$

where y_n is the normalized value of the item of the sequence;
Y_i is the current not normalized value of the item of the sequence;
L is the number of elements of the sequence «science» in a particular geolocation group.

The normalized geolocation distributions by clusters of the first group are presented in Figure 4. Analysis of the conformity of the frequency of references of the messages about science and the scientific areas of one or another cluster was conducted by each group of countries and by each cluster. Thus, it was concluded that in any of one or another group of society, topic about science is especially mentioned in the research of a particular cluster.

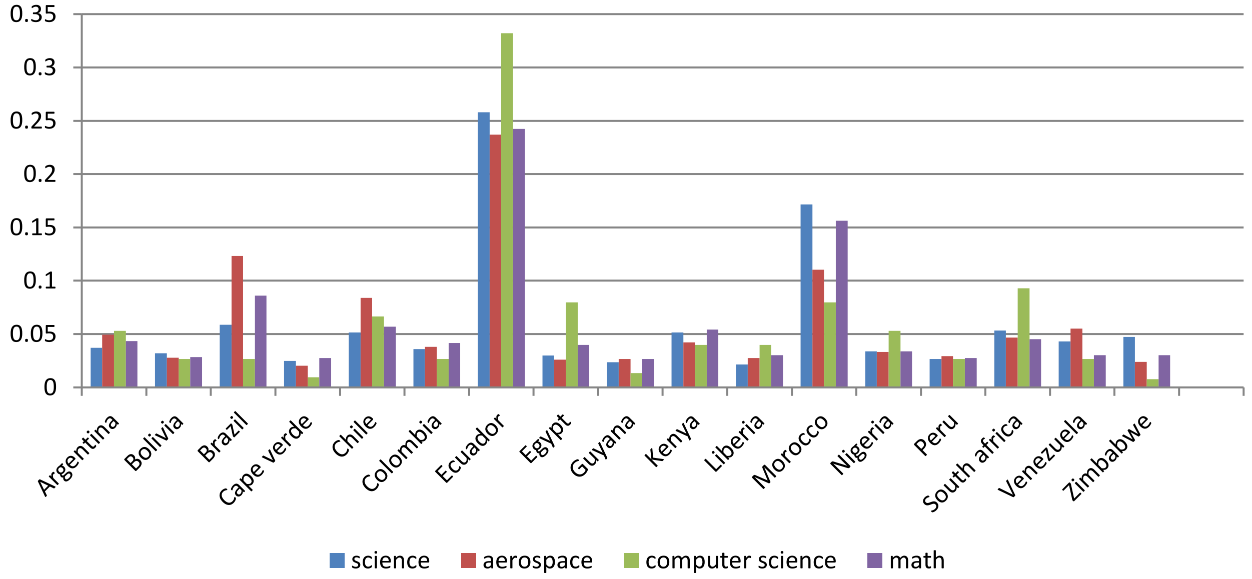

Figure 4: The normalized geolocation distribution by clusters of the first group of countries

For more austerity experiment, sequence distributions of tweets by groups of countries in different clusters were compared with the sequence of distributions of tweets about science for the same group of countries, using count values of the coefficient of correlation of Pearson [11] and checks using the Kolmogorov-Smirnov [12]. The results are shown in Table 1 and 2.

IV. DISCUSSION

The results suggest that the users of the social network of Twitter primarily mentioned the topics of science, which are connected to the space, a little less often with information technology and even more rarely with mathematics, medicine and physics. A more detailed analysis via methods of mathematical statistics shows that there was high dependence of the distributions of using messages, which contain three themes of messages, which contain references to a science by locations. Based on these findings, we can conclude that indeed there is a clear relationship between the spread of the topics, which are discussed in social networks, defined by clusters «space», «computer science», «fundamental sciences» and «science» as such. In other words, the more users talk about the terms of the three clusters in a given geographical location, the more they talk about the science and the most likely interested in it. Moreover, in the group of developed countries (Europe, USA, Canada, Australia), the distribution of the concept of «science» is related with the concepts of all the three clusters (Ys, Yc, Yf), but most of all – the space theme. In the group of countries, specifically in Asia and Russia, the ratio of these two clusters with science is expressed not as clearly as in the previous location. However, the relationship between the distributions are also high, and most of all pertaining to the topics of information technology and fundamental sciences (math, physics, medicine). For the group of developing countries (Africa, South America), the relationships between distributions, namely Ys, Yc, Yf, and X are even weaker. However, there was less for the fundamental sciences group: For the users of social networks in this region, science is associated primarily with mathematics and physics. These findings can be explained by the fact that in the countries of Africa and South America, information and space technologies are not as well developed as in North America and Europe; in Asia information technology is more widespread than space; and in the countries of the third group the popularity of space programs and technologies in the United States, a country with the largest amount of data generated in the social network Twitter, plays an important part.

V. CONCLUSIONS

Due to its huge popularity in the whole world, social networks contain a large amount of data. One can get a slice of public opinion on various topics and issues of concern when extracting information from them. The collection and processing of this information makes it possible to use methods and tools of Big Data. Wherein, firstly it carried out continuous work with the data in real time followed by clustering the data by geographical locations. Then, further analysis using methods of mathematical statistics, within these locations was carried out. Similarly, it was found that users of the social network Twitter, refer to the topic of science in the social networks primarily have in mind all that is connected with the space, a little less often with information technology and even more rarely with mathematics, medicine and physics.

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.

Table 1
The value of the Pearson correlation coefficient for clustered sequences of three groups of countries

Clustered sequences:	The value of the Pearson correlation by the 1 group of countries	The value of the Pearson correlation by the 2 group of countries	The value of the Pearson correlation by the 3 group of countries
science and space	0,99802679	0,817716579	0,837834028
science and computer science	0,904232913	0,876366191	0,71120487
science and fundamental sciences	0,891262183	0,845892724	0,93754061

Table 2
The level of significance by the Kolmogorov table for clustered sequences of three groups of countries

Clustered sequences:	The level of significance by the Kolmogorov table for the 1 group of countries	The level of significance by the Kolmogorov table for the 2 group of countries	The level of significance by the Kolmogorov table for the 3 group of countries
science and space	1,0	0,76	0,8

REFERENCES

[1] A., El-Hoiydi, and J.D., Decotignie. “WiseMAC: an ultra-low power MAC protocol for the downlink of infrastructure wireless sensor networks,” in Ninth International Symposium on Computers and Communications, *Proceedings. ISCC*, 2004, vol. 1, pp. 244–251.

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

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

[4] 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*, 2015. Pp. 179-184.

[5] 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", *Proceedings of the ACL System Demonstrations. Association for Computational Linguistics,* 2012,. 115-120.

[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] D., Borthakur, J., Gray, J., Sarma, K., Muthukkaruppan, N., Spiegelberg, H., Kuang, and R., Schimdt. "Apache Hadoop goes realtime at Facebook", *Proceedings of the ACM SIGMOD International Conference on Management of Data*., 2011, 1071-1080.

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

[10] S., Wanderman-Milne, N., Li. "Runtime Code Generation in Cloudera Impala", *IEEE Data Eng*. Bull 2014, T. 37, No. 1. 31-37.

[11] J., Dean, S., Ghemawat. "MapReduce: simplified data processing on large clusters" *Communications of the ACM*. 2008. T. 51. No. 1. pp. 107-113.

[12] V. V., Rykov, B. Yu Itkin, "Mathematical statistics and design of experiment", Moscow: Russian State University of Oil and Gas named by I.M. Gubkin, 2008.

[13] Kolmogorov-Smirnov test [Electronic resource], Academician. Mathematical encyclopedia, URL: http://dic.academic.ru/dic.nsf/enc_mathematics/2279/ (accessed: 26.05.2015).