CLASSIFICATION OF TEXT DATA FROM THE SOCIAL NETWORK TWITTER

I.A. Rytsarev, A.V. Blagov

Samara National Research University, Samara, Russia

Abstract. Social networks play an important role in the modern world, and it is important to define the important and popular topics discussed. This article deals with data collection from the social network Twitter, and further clustering and classification of the collected data.

Keywords: bigdata, data processing, data analysis, clustering, classification, tf-idf, latent dirichlet allocation.

Citation: Rytsarev IA, Blagov AV. Classification of Text Data from the Social Network Twitter. CEUR Workshop Proceedings, 2016; 1638: 851-856. DOI: 10.18287/1613-0073-2016-1638-851-856

Introduction

The extra-large volumes of data in information technology - data sets the size of which is beyond the capabilities of typical database (DB) for collecting, storage, management and analysis of information [1].There are many series of approaches, tools and methods of processing such extra large volumes of structured and unstructured data[1-4].

The concept of big data means working with the vast volume of information and varied composition, very frequently updated and located in different sources in order to increase efficiency and create new ones.

At the moment, social networks are at the peak of their popularity, millions of people are already use Facebook and Twitter. Many companies need to analyze the data collected from social networks to assess the relationship of users to their products [5]. Also the analysis of this area is used in solving security issues [6]. Having collected and clustered text data from the social network, it is possible to identify the main themes and events discussed by social network users in different cities and countries.

1 The clustering of text information based on the frequency analysis

Clustering (or cluster analysis) - is the task of partitioning a set of objects into groups called clusters. Within each group must be "similar" objects, and objects of different groups should be as different as possible, at the same time certain measure has to be determined. The main difference between clustering and classification is that a lists of groups are not clearly defined and determined during the process of the algorithm. The main goal of clustering is the searching of existing structures [7].
One of the main methods of frequency analysis is to count the number of occurrences of each word in document. Based on the information received you can create so-called "tag cloud" - a visual representation of the weight of the words in the document [8].

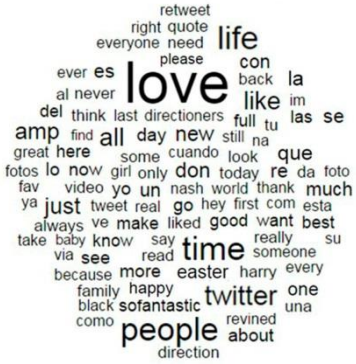

Fig. 1. "Tagcloud" of the English-speaking segment of the social network Twitter

However, as a result of the data processing at the output we obtain that different articles, prepositions and other service parts of speech will have the largest number of occurrences. Therefore, for the most honest evaluation of the meaning of the word is necessary to use measures that will not only count the number of occurrences of the word in the document, but also take into account the number of occurrences of the word in other documents. An example of such measures is the TF-IDF. During the calculation, the weight of a word is proportional to the amount of use of the word in the document, and is inversely proportional to the frequency of use of the word in other documents of the collection [9]. Thus, TF-IDF measure is the product of two factors:

$$tfidf(t,d,D) = tf(t,d)\, idf(t,D), \tag{1}$$

where $tf(t,d) = \dfrac{n_i}{\sum_k n_k}$ and $idf(t,D) = \log \dfrac{|D|}{|(d_i \supseteq t_i)|}$

After determination by the TF-IDF measure, clustering can be performed by various algorithms, such as k-means, for example [14].

2 Classification of the text information based on the approach with machine learning

Also, there is another approach to solve the problem of the classification - the classification of the information through the machine learning.
Machine learning - the process through which the machine (computer) is able to show the behavior which was not explicitly programmed. There are two types of learning: inductive and deductive.
In the work of researchers involved in cluster analysis of the text information in various search engines, the inductive measure Word2vec is frequently used [10-11].
The principle of the measures is to find connections between the context, the word according to the assumption that the words that are in similar contexts, tend to mean similar things, i.e., be semantically close. More formally, the task is: to maximize the cosine proximity between vectors of the words (scalar product of the vectors) that appear next to each other, and minimizing the cosine proximity between the vectors of words that do not appear next to each other. “Next to each other” in this case means “in close contexts”.
Word2vec analyzes the use of words contexts and concludes that they are or are not close in meaning. Since word2vec making such conclusions based on large amounts of text, the conclusions are quite adequate The algorithms on which word2vec is base dare described in detail in [12-13].
The examples of vector distances obtained by word2vec are in Table 1.

Table 1. The vector distances between the word «France» and other words by word2vec measure

The word	The vector distance
Paris	0.978443
Spain	0.665923
Belgium	0.665923
Netherlands	0.652428
Italy	0.633130
Portugal	0.577154
Russia	0.571507
Germany	0.563291

One type of a deductive approach can be considered is the Latent Dirichlet Allocation (LDA). This generative model that allows to explain the results of observations with the help of implicit groups, that allows to receive an explanation of why some of the pars of data are similar. Typically, when using this approach, you identify a limited number of topics and further states that each document is a mixture of a small number of topics [15].
For more detailed analysis, it is best to combine different approaches and techniques depending on the amount of processed data.

3 The method of collecting information

To research the work of TF-IDF algorithm software tool that allows to collect data directly from the Twitter servers has been developed. The implementation is based on an open Twitter API 2.0 interface. As the object of the study the tweets from Samara region were taken, for this as a selection criterion of the messages geolocation was set to the Samara region (including all settlements of the region). All tweets collected in such a way need to be clustered using TF-IDF metric for short messages of 140 characters in length and the algorithm k-means.

To carry out the data collection from Twitter server a request containing a consumer key and consumersecret key was sent. In reply we have oauth.accessToken, oauth.accessTokenSecret which give us the ability to retrieve data from servers.

The second step is to send the query-request, in response to which a set of tweets returns.

Next, the third step is the counting the TF-IDF metric values for each message.

4 Results

The data was collected for 24 hours, by the query-request, which characterizes the Samara region. As the result over 6000 messages has been collected. By applying metrics TF-IDF and the k-means algorithm 22 clusters were obtained. On example, one of the obtained clusters (Figure 2) shows that the messages are similar in meaning, but among them there are messages with "foreign" theme.

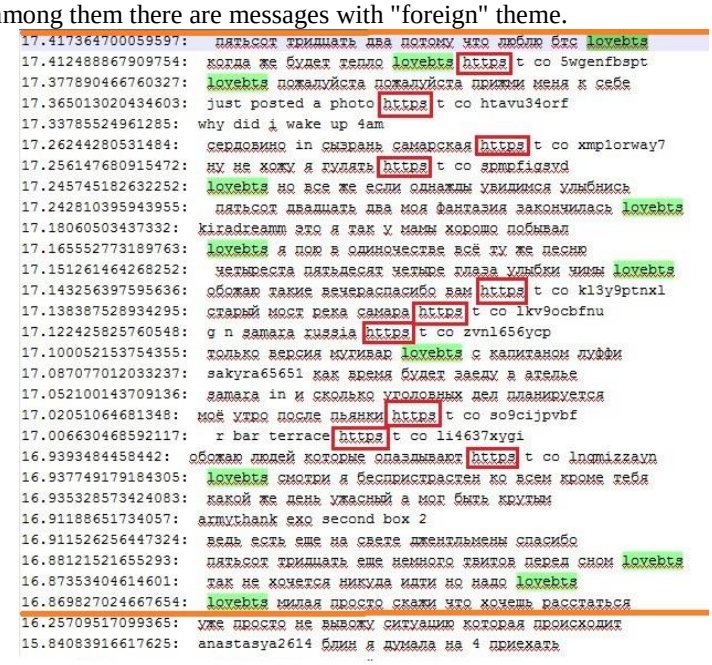

Fig. 2. An example of one of the clusters obtained

Apparently, such is not quite accurate result was obtained due to the fact that studied Twits have a 140 character limit. For this reason, for more accurate clustering and further classification it is necessary to modernize TF-IDF measure, introducing additional weighting coefficients corresponding to the number of symbols (words) in the message.
Moreover, high density of the clusters (Figure 3) shows the need for revision of the metric.

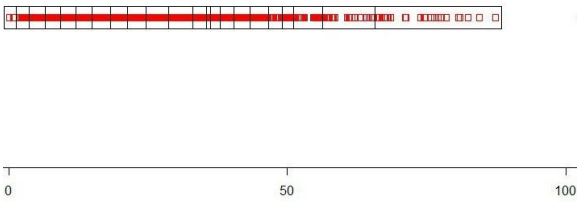

Fig. 3. The distribution of TF-IDF metric values of processed data on the number line

TF-IDF metric works for short messages with relative accuracy. For this reason it is necessary to optimize this metric: adding weight coefficients, the coefficients associated with the hashtag, the introduction of the normalization coefficient associated with the length of the message and the number of words in it, etc.

Conclusion

Issues related to clustering and further classification of text data are relevant in relation to the enormous spread of social networks and online services worldwide. Approaches and techniques presented in the article are planned for testing on text data collected from Twitter social network in the Russian segment. Collecting the necessary data is being produced by means of the developed software system, based on time zones and geolocation. It is planned to develop the subject in the direction of the output and optimization of parallel clustering algorithms.

References

1. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM, 2008; 51: 107-113.
2. Vossen G. Big Data as the new enabler in business and other intelligence. Vietnam Journal of Computer Science, 2014; 1: 3-14.
3. Tamhane DS, Sayyad SN. Big Data Analysis Using Hace Theorem. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2015; 4: 18-23.
4. Kazanskiy NL, Protsenko VI, Serafimovich PG. Comparison of system performance for streaming data analysis in image processing tasks by sliding window. Computer Optics, 2014; 38(4): 804-810.

5. Tan W, Blake M, Saleh I, Dustdar S. Social-network-sourced big data analytics. IEEE Internet Computing, 2013; 5: 62-69.
6. How “Big Data” help to improve security. URL: http://www.computerra.ru/108760 /security-n-big-data/.
7. Data Mining tasks. Classification and cauterization [In Russian]. URL: http://www.intuit.ru/ studies/courses/6/6/lecture/166.
8. Blagov A, Rytcarev I, Strelkov K, Khotilin M. Big Data Instruments for Social Media Analysis. Proceedings of the 5th International Workshop on Computer Science and Engineering, 2015; 179-184.
9. Ramos J. Using tf-idf to determine word relevance in document queries. URL: https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf**.**
10. Wang H. Introduction to Word2vec and its application to find predominant word senses. URL:http://compling.hss.ntu.edu.sg/courses/hg7017/pdf/word2vec%20and%20its%20application%20to%20wsd.pdf>.
11. Yu M, Dredze M. Improving lexical embeddings with semantic knowledge. Association for Computational Linguistics (ACL), 2013; 2: 545-550.
12. Mikolov T, Chen K, Corrado G, Dean J**.** Efficient Estimation of Word Representations in Vector Space. URL: http://arxiv.org/pdf/1301.3781.pdf.
13. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. Advances in neural information processing systems, 2013; 3111-3119.
14. MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability, 1967; 281-297.
15. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. The Journal of machine Learning research, 2003; 3: 993-1022.