Development and implementation of social network data collection services to improve the human environment

I A Rytsarev[1], A V Blagov[1] and M I Khotilin[1]

[1]Samara National Research University, Moskovskoe Shosse 34, Samara, Russia, 443086

Abstract. The article discusses the need to develop and implement services for the collection of social networking data for the detection of environmental problems. The authors offer methods and tools for collecting and analyzing data on identified problems from social networks.

1. Introduction

There are many ways of monitoring a person's environment, such as photography from satellites or unmanned aerial vehicles, equipped with cameras, transportation, or installation of stationary cameras, etc. Different methods of approach can be used to detect the necessary objects [1-2]. Often, this problem is solved by the forces of certain municipal services. As a rule, this approach is associated with high labor costs. Such information services as thematic Internet sites, for example, are becoming popular: https://rosyama.ru/, http://moyasamara.com/ etc. It is worth noting that to apply through these sites, you must at least remember the name of the resource, its address. Moreover, creating an application, the author must fill out information about yourself. Among some people, these restrictions can create some difficulties. At the same time, social networks are becoming increasingly popular [3]. 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. 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 [3-4]. 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.

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 and cities can be received. It helps to estimate the popularity of the selected theme in specific geographical locations.

Many people have smartphones and other personal mobile gadgets, equipped with the ability to access social networks. Due to this, each user generates a large amount of data that may be of interest for different areas and areas of activity. A lot of research works are devoted to the collection, processing and analysis of social network data [4-6]. It is worth noting that for the user of social networks themselves networks can be the most convenient resource for posting information, including on various environmental problems: landfills, accidents, pits, fires, etc.

This article describes the implemented service for the operational collection of information about the problems of the environment in one of the most popular social networks Twitter. The service allows you to collect the necessary information on-line for any necessary geolocation.

2. Data collection algorithm, their processing and classification

There are many social networks (YouTube, Twitter, Facebook, VK, Instagram, etc.). With their help, users exchange various information (text information, video images, etc.). The social network Twitter was chosen for the research. This was done for the following reasons:

- users of this social network mainly share text information which is easier to process;
- the network provides open access to its data (there is no restriction on access to server data streams);
- twitter is the second most popular social network (after Facebook, which does not provide open access to its data) among users worldwide;
- twitteris not a substantive network and most broadly reflects public opinion on issues of interest to us.

The task of collecting the necessary information from social networks on environmental problems can be divided into data collection, filtering, processing and, if necessary, classification. The data generated on-line are of the greatest value for obtaining information on various problems of the habitat. In this case, the filtering can go on a certain, interesting geolocation, and on the subject of the published content.

The following information is generally required when collecting Twitter data:

- idofthemessage;
- messagetext;
- listofmessagehashtags;
- geolocation;
- date and time of the message.

The data collection algorithm developed in the framework of the study, in addition to the indicated filtering, uses an additional element of user feedback through an automatic request for clarification of information about the published message. Figure 1 shows the scheme of the algorithm for collecting the necessary data from the social network.

The social network data generated in on-line mode is collected using the F1 filter for the specified geolocation. Next, using the content filter F2, configured by key words or topics, the necessary data on habitat problems are collected. The data obtained by double filtering can be schematically divided into two parts: the data about the author-A and the message-M. In this case, according to the developed algorithm, the author of the message automatically receives a clarifying request Q, to provide additional information, such as the exact address or coordinates of the problem described in the message M, or its nature. The author a, receiving this request, leaves message N, containing the necessary information. We can say that according to the developed algorithm gathers the following data set: $\sum_{i=1}^{k}(M_{gi} + N_{gi})$, for all k users A_i, the message authors M_{gi}, generated by geolocation g, and send the additional information N_{gi}.

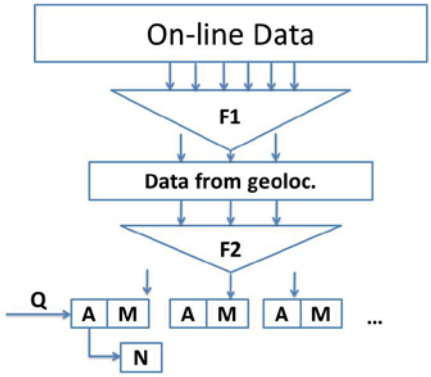

Figure 1. The general scheme of the algorithm of collecting social network data.

After collecting the required data is often required to classify them by reference to certain issues: landfills, pits, accidents, etc. To solve this problem, one can use the key word "hashtag." this message [7], or other approaches, for example N-grams [8]. For a more detailed analysis, the algorithm of collective decision-making developed in the framework of this study can be used.

The method of collective decision - making is based on the uniqueness of each word. It is based on the dictionary, which stores the words and their categories to which they belong.
The dictionary is formed with the participation of an expert who teaches the system, and has the ability to make changes to the dictionary to clarify it.

Any collected post each word from the text message $T = t_1 t_2 t_3 ... t_n$ makes to a dictionary $U = u_1 u_2 u_3 ... u_m$ to obtain a list of categories:

$$\delta(u_i) = \{\underbrace{k_j ... k_h}_{z}\},$$

where $\delta(u_i)$ - is the function of extracting categories of the word u_i from the dictionary; $\underbrace{\{k_j ... k_h\}}_{z}$ - the vector of z elements k (element list categories $K = k_1 k_2 k_3 ... k_l$), to which the word u_i belongs; $1 \le j, h, z \le l$.

If $\gamma(t_i) = u_i$ ($\gamma(t_i)$ - the function of the stemming text, then $\delta(t_i) = \delta(u_j) = \overline{k}$. If there is no word in the dictionary, it is given to the expert for classification with subsequent addition to the dictionary.

A distinctive feature of this algorithm of collective decision-making is the presence of a dynamic threshold value: when determining the threshold value, the number and weight of the categories of words mentioned in the text are taken into account.

As a result of text processing we have a list of categories (which include words) on the basis of which the algorithm calculates the threshold value of the "weight" of the subject and classifies the source text.

3. Results and discussions

The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in table 2 should be used.

Data collection on Twitter can be carried out through Apache Ambari and Flume software products, this method is described in more detail in [9]. However, to collect data using a number of filters, it is often more convenient to develop your software product using standard libraries (twitter4j, tweepy, etc.) [10].

Within the framework of this study, a software product was developed in the Python programming language, containing an authorization module, a data collection module and a filtering module. This software allows you to collect data on geolocation, keywords, user, and cache all media files of the

user. To avoid interruptions in the operation of the software product associated with exceeding the limits set by the social network Twitter, a lot of authorization keys are sewn into the software product. The software works in real-time monitoring mode, and can also make requests for information lying on servers.

Launches of the developed software product were made on the basis of Launches of the software complex were made on the hardware and software complex of data processing of the very large volume of the laboratory for data processing of the very large volume of the Samara University. Hard ware and soft ware complex consists of:

- the software and hardware complex of storage and analytical analysis of IBM Puredata for Analytics (Netezza) structured data with a volume of disk space of at least 96Terabyte (taking into account 4-times compression and full data replication);
- Hadoop-clusterраспределённогохраненияианалитическойобработкинеструктурированныхданных (IBM x3630 M4 management server (two Intel Xeon Processor E5-2450v2; 96 GB of memory; 2 600GB disks) and four IBM x3630 M4 data processing servers (two Intel Xeon Processor E5 2450v2 processors; 96 GB of RAM; 8Terrabyte of disk memory)).

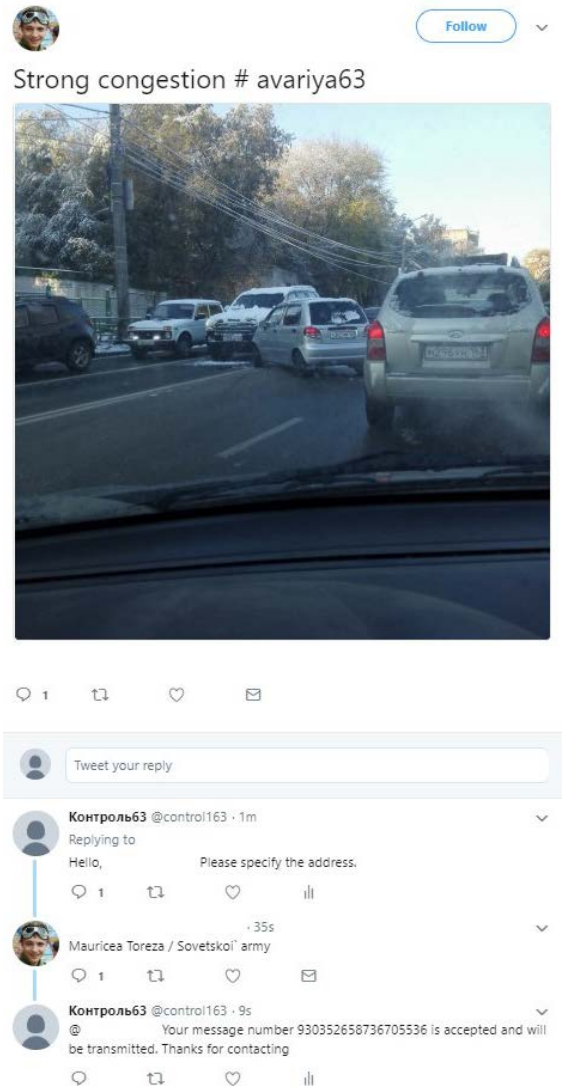

Figure 2. The receive notification and clarifying the question on Twitter (example 1).

The hardware and software complex has access to the world wide web at a speed of 10 GB per second, which allows you to work with a large flow of messages.

For each user of the social network Twitter implemented interface for interaction is a bot account @contriol63. The geolocation filter is set to the sixty-third region of the Russian Federation, which is the Samara region.

A user who is going to write in his message about the problem of the environment, can add to your message or the name of the bot account, or its address, or a set of keywords presented in the table №1.

Table 1. Keywords necessary for use in the message.

Problem	Keywords
the unauthorized dump	#dump63, dump63, #trash63, trash63, #svalka63, svalka63, #musor63, musor63,
the fire	#fire63, fire63, #pozhar63, pozhar63,
pitsonroads	#pit63, pit63, #yama63, yama63
breakdown	#breakdown63, breakdown63, #avariya63, avariya63
danger	#danger63, danger63, #opasnost63, opasnost63,
disturb	#disturb63, disturb63, #narushenie63, #narushenie63,
no designation name address of the bot	andcontrol63, @control63, контроль63, @контроль63

The list in table № 1 may be supplemented.

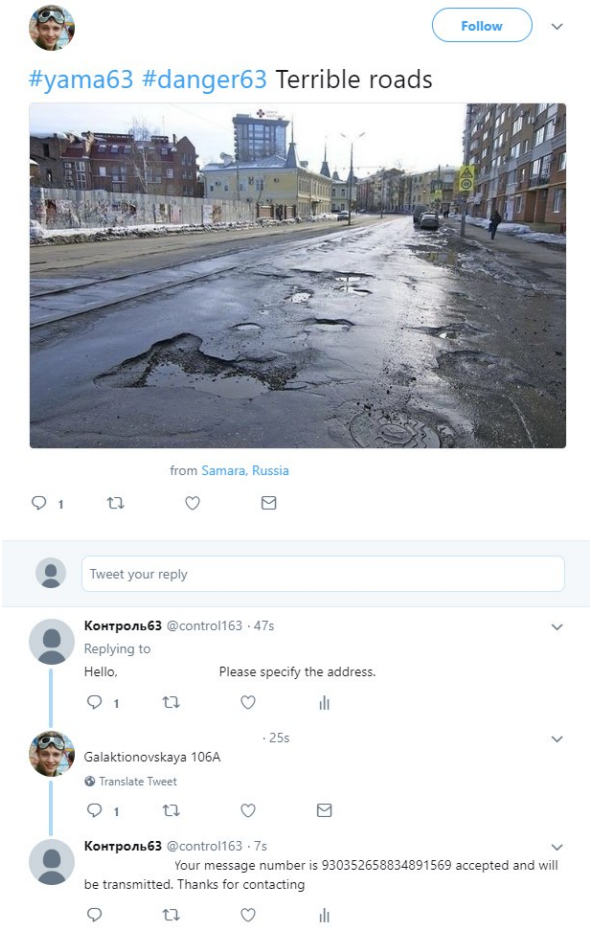

Figure 3. The receive notification and clarifying the question on Twitter (example 2).

A written message containing any of the keywords in table 1, with geolocation by 63 region, is collected and classified by type. The author of the message is automatically notified of the collection and asked a clarifying question (figure 2, figure 3).

The collected pool of reports on various problems of the human environment is valuable information and can serve for the rapid targeted elimination of accidents and violations. At the same time, the service is very convenient for all users who simply continue to use their social networks, putting there information about the problems that have arisen. It is also worth noting that the developed service can be of practical value only if it is supported by structures and services dealing with the problems of the human environment. The main motivation for the users who post information on this topic, can only be a rapid response and subsequent problem solving. At the same time, in case of successful testing and implementation within the Samara region, the project can be scaled to other regions and even countries.

4. Conclusion

The result of the work is the development and implementation of the service to collect the necessary information about accidents and violations in the social network Twitter for a certain geolocation. Additionally, a classification tool for collected messages is also implemented.

In the future, the developed service can be: first, scaled to other social networks, for example, Vkontakte, and synchronized between them, and secondly, it can be supplemented with an image processor (photos attached to the message) and video.

5. References

[1] Epifancev B N, Pyatkov A A and Kopeykin S A 2016 Multi-sensor systems for monitoring access to restricted areas: capabilities of the intrusion detection video analytical channel *Computer Optics* **40(1)** 121-129 DOI: 10.18287/2412-6179-2016-40-1-121-129

[2] Vizilter Y V, Gorbatsevich V S, Vishnyakov B V and Sidyakin S V 2017 Object detection in images using morphlet descriptions *Computer Optics* **41(3)** 406-411 DOI: 10.18287/2412-6179-2017-41-3-406-411

[3] Tan W, Blake M, Saleh I and Dustdar S 2013 Social-network-sourced big data analytics *IEEE Internet Computing* **5** 62-69

[4] 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

[5] Xu X 2007 Scan: a structural clustering algorithm for networks *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* 824-833

[6] 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

[7] Krokos E , Samet H and Sankaranarayanan J 2014 A look into twitter hashtag discovery and generation *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* 49-56

[8] Mikhaylov D V, Kozlov A P and Emelyanov G M 2017 An approach based on analysis of n-grams on links of words to extract the knowledge and relevant linguistic means on subject-oriented text sets *Computer Optics* **41(3)** 461-471 DOI: 10.18287/2412-6179-2017-41-3-461-471

[9] Rytsarev I A and Blagov A V 2015 Construction of activity models of users of social networks *Proceedings of the Information Technology and Nanotechnology* (Samara: Samara National Research University) 216-220

[10] 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

Acknowledgments

The work has been performed with partial financial support from the Ministry of Education and Sciences of the Russian Federation within the framework of implementation of the Program for Improving the Samara University Competitiveness among the World's Leading Research and Educational Centers for the Period of 2013-2020s.