Analysis of Social Networks Using an Information and Analytical System

Rodion Filippov*, *Yuriy* Leonov, *Aleksandr* Kuzmenko, and *Timofey* Shestakov

Department of Computer Technologies and Systems, Bryansk State Technical University, 7 Bulvar 50- letiya Oktyabrya, Bryansk 241035, Russian Federation

Abstract. The subject of the study is the analysis of social networks and the construction of an information and analytical system to automate data monitoring and mining. Modern social network analysis systems are reviewed, and the distinguishing features of these systems are given. Various methods of social network analysis and tasks that can be solved using these methods are described. The effectiveness of the methods to determine the text sentiment is compared.

1 Introduction

Thanks to the rapid development of social networks, large amounts of users' personal data such as comments, photos, videos, audio information, geotags, and others have become publicly available. This opens up great opportunities for solving research and business problems that are difficult to solve effectively without a great deal of data.

Specialists from all over the world use social network data to create and model social, economic and other processes aimed at solving state tasks in order to create tools for influencing these systems, as well as for creating analytical systems and business applications.

Developing such systems has a number of problems and features that need to be solved. The first difficulty in creating such systems is large amounts of data, which is both an advantage and a disadvantage [12]. Large amounts allow to get more accurate research results, but they require the construction of a complex distributed system architecture that contributes to increase the efficiency of the system proportionally to the added computing power. The second problem is that processing and storage of social network data requires the development of special algorithms that allow to take into account the features of the subject area as well as the infrastructure solutions. There are other problems related to data privacy, restricted access to data, and weak data structure.

The analysis of social networks is understood as the solution of such tasks as defining the text sentiment, determining the target users, searching for associative rules, calculating the performance indicators of the community and data visualization.

* Corresponding author: libv88@mail.ru

2 Analysis of previous paper results

There are various foreign and domestic systems for monitoring and analyzing data in social networks.

Social network analysis systems can be classified according to the following criteria:

1. Data analysis methods. There are basically two classes of methods used to analyze social networks: statistic analysis (SA) and graph analysis methods. Classification methods are used for semantic analysis of the text and text sentiment (TS), statistic and clustering methods are used to determine the target users, visual analysis (VA) is used to demonstrate the obtained data and restrictions. It is also often possible to search by keywords (PK) for subsequent analysis of related content. Retrospective analysis (RA) allows to consider the dynamics of objects changes.

2. Analysis objects. The system can analyse various social network objects: informational messages, opinions, subnets and communities, individual users, and external nodes.

3. Set of data sources. The more data sources a system has, the more accurate the results of research using technologies such as BigData and deep neural networks can be.

4. System users. Depending on the target users of the system, analysis methods and objects may differ. Also, it is important for commercial organizations (CO) to have API of the system, that is the ability to upload reports. To use these systems in government organizations (GO), it is necessary to meet certain standards and be included in the unified register of Russian software. For scientific and educational institutions (EI), an important factor is the ability to use the systems for scientific purposes and the availability of good documents and the pricing policy of the company.

5. Characteristics. Each system has additional characteristics that distinguish them from their competitors [1,11]. Among the most popular systems are: Brand Analytics, IQBuzz, Agorapulse, SemanticForce, Talkwalker. Each system has its own characteristics and it works in a specific area of analytics and data collection (Table 1).

Table 1. Comparative table of social network analytical systems

System / Tag	Brand Analytics	IQBuzz	Agorapulse	Semantic Force	Talkwalker
Users	CO, GO	CO	CO	CO, GO, EI	CO
Analysis methods	TS, SA, PK	TS, SA, PK, RA, VA	SA, VA	TS, SA, VA	TS, SA, VA, RA
Analysis objects	Information messages	Information messages, communities, users, external nodes	Information messages, opinions, communities	Information messages, opinions, communities	Information messages, opinions, communities, images
Data sources	VK, Facebook, OK, Instagram, YouTube, Telegram, mass media	LiveJournal, VK, YouTube, Instagram, Twitter	Facebook, Twitter, LinkedIn, Google+, Instagram	Facebook, Twitter, VK, OK, YouTube	Facebook, Twitter, LinkedIn, Google+, Instagram
Characteristics	67 languages support, unloading of reports	API, retrospective review up to 10 years, unloading of reports	CRM for audience segmentation, scheduled posting	API, integration with Google Analytics, text rubrication	187 languages support, unloading of reports

As you can see in Table 1, most systems are designed to work with commercial organizations. The systems that are included in the unified register of Russian software also work with governmental organizations and scientific institutions.

In the systems considered, the following analysis methods are often used: text sentiment analysis, statistical and visual analysis. Innovative systems implement analysis methods using search, keyword analysis, retrospective analysis, and image analysis. The most popular objects of analysis are informational messages, opinions, and communities.

The choice of data sources depends on the regional location of the company's target audience and the scope of analytical activities. Facebook, Twitter, Instagram, and Google + are analyzed by the systems designed for America and Europe, while the domestic systems pay special attention to VK, OK, YouTube, Instagram, and mass media.

3 Research materials and methods

The research material includes social networks, that is information messages, opinions, subnets and communities, individual users, and external nodes.

Data on communities (*Community*):

$$Community=\{ CCM, CC, CD, CP, CS\}, CC \in Categories, CP \in Posts, \quad (1)$$

where *CCM* (count of community members) is the number of community members –, *CC* is community category, *CD* is community description, *CP* stands for community posts, *CS* means community subscribers.

Users are described by the following data set:

$$User=\{UA, US, UCo, UCi, UCF, UI, UE\}, UCO \in Countries,$$

$$UCi \in Cities, US=\{Male, Female\}, UI \subset Categories, \quad (2)$$

where *UA* is user age, *US* is user sex, *UCo* is user country, *UCi* is user city, *UCF* (count of friends) is the number of friends, *UI* stands for user interests, *UE* means user education.

Information on posts:

$$Post = \{PD, PV, PL, PC, PR\} \quad (3)$$

where *PD* is post description, *PV* (post views) is the number of post views, *PL* stands for post likes, *PC* is the number of post comments, *PR* means post reposts.

Data on comments:

$$Comment = \{CT, CL, CV\} \quad (4)$$

where *CT* is comment text, *CL* (comment likes) is the number of comment likes, *CV* means the number of comment views.

Data analysis is performed using the following methods: Data Mining, Statistical Analysis, Visual Analysis, and Retrospective Analysis.

3.1 Data Mining

Data mining is the process of discovering previously unknown, non-trivial, practically useful and available to interpretation knowledge in raw data, which is necessary for making decisions in various human activities [2,13]. Data Mining methods were used to solve the following tasks: to determine the audience's reaction to the content, to define the target profile of subscribers, and to find the dependencies between user interests.

Determination of the audience's reaction to the content

One of the tasks of data mining in social networks is to determine the audience's reaction to the content. This problem can be solved by using methods of defining text sentiment, by analyzing comments to the content. Solving this problem will help determine the mood of the audience and will be useful when choosing a group for advertising products or services [3].

This problem can be solved by using two classes of methods: methods based on defining text sentiment from pre-compiled tone dictionaries, and machine learning methods such as Bayes classifier, *k*-nearest neighbour method, and support vector machine.

1. *Tone dictionaries*. In the method of tone dictionaries, the text can be estimated by a scale containing the amount of negative and positive words based on the amount of emotive vocabulary found.

2. *Bayes classifier (Naive Bayes)* is a broad class of classification algorithms based on the principle of maximum a posteriori probability.

3. *k-nearest neighbour method (KNN)* is a metric algorithm for automatic object classification or regression.

4. *Support vector machine (SVM)* is a set of similar learning algorithms with a teacher that are used for classification and regression analysis tasks [4,14].

During the research the most accurate method was chosen to determine the text sentiment.

Definition of the target profile of community subscribers

To determine the target profile of users in a certain community, the method of cluster analysis was used.

K-means method was applied for clustering. *K*-means method is a cluster analysis method that aims to divide n observations (from space) into k clusters.

K-means algorithm splits set x into k sets $S_1, S_2, ..., S_k$, in such a way as to minimize the sum of squared distances from each point of the cluster to its center (the cluster center). We introduce the following notation, $S = \{S_1, S_2, ..., S_k\}$. Then the action of *k*-means algorithm is equivalent to minimizing the total square deviation of the cluster points from the centers of these clusters:

$$V = min \sum_{i=1}^{k} \sum_{X \in S_i} \rho(X, \mu_i)^2 \quad (5)$$

where μ_i is the center *i*-cluster, $i=1,k, \rho(X, \mu_i)$ is a function of the distance between *x* and μ_i.

Euclidean distance can be used as a distance function:

$$\rho(X, \mu) = \sqrt{\sum_{i=1}^{d} (x_i - \mu_i)^2}, \quad (6)$$

where $X=\{x_1,x_2,...,x_k\}$, $x \in R^d$, d is X vector dimension.

The advantages of the algorithm are relatively high efficiency along with simple implementation, high quality clustering, and the ability to parallelize the algorithm [5].

The analyst enters the number of clusters into which the community audience should be divided, then the most common user characteristics are determined in each cluster, such as: age, gender, city, interests, and others.

Definition of dependencies between user interests

Frequency analysis of element sets and the study of association rules can be used to find patterns between user interests. One of the most popular algorithms for finding associative rules is Apriori.

Apriori is an algorithm for the frequency analysis of element sets and the study of association rules in relational databases.

Given $I=\{i_1,i_1,...,i_n\}$ – a set of user characteristics called elements, $D=\{t_1,t_2,...,t_n\}$ – a set of transactions, where each transaction has a unique identifier and $D \subseteq I$.

An association rule is an implication of the following form:

$$X \rightarrow Y, \text{ где } X, Y \subseteq I. \quad (7)$$

To select a rule from a set of all possible rules, constraints on various measures of significance are used. The most well-known constraints are the minimum support threshold and the minimum fidelity threshold.

The rule support shows the frequency with which set $X \rightarrow Y$ occurs in the set of transactions [6]. Support for set X related to D is defined as the ratio of the number of transactions t in the database containing set X to the total number of transactions:

$$S(X \rightarrow Y) = \frac{\left|\{t \in D; X \subseteq t\}\right|}{|D|}. \quad (8)$$

The fidelity of the rule shows the frequency with which $X \rightarrow Y$ is observed in the data set [7]. The fidelity value of the rule related to a set of transactions D is the ratio of the number of transactions that contain both set X and set Y to the number of transactions that contain set X:

$$C(X \rightarrow Y) = \frac{S(X \cup Y)}{S(X)}. \quad (9)$$

3.2 Statistical Analysis

The effectiveness of social media marketing depends on the values of key performance indicators (KPIs) in advertising groups.

All metrics can be divided into the following categories:

Metrics for evaluating the dynamics of subscribers

1. The number of subscriptions for a period (Follows).
2. The number of unsubscribes for a period (Unfollows).
3. The number of views (Views) – as a rule, the total indicator for all community records for a period is used:

$$Views = \sum_{i=1}^{n} PV_i, \quad PV_i \in CPP, CPP \subset CP, \quad (10)$$

where n is a number of posts, PV_i is a number of *i-post* views, CPP is a set of posts for a certain period.

4. Reach shows the number of users who at least once have contacts with the community posts:

$$Reach = \sum_{i=1}^{n} \left(PL_i + PR_i + PC_i\right), PL_i, PR_i, PC_i \in CP \quad (11)$$

where n is the number of posts, PL_i is the number of likes in *i*-post, PR_i is the number of reposts of *i*-post, PC_i is the number of comments for *i*-post.

Metrics for evaluating audience feedback

Metrics that reflect the user's reaction to the content. The most well-known metrics are likes, comments, and reposts.

1. Love Rate (*LR*) is an average number of likes in terms of audience size:

$$LR = \frac{1}{n} \sum_{i=1}^{n} \frac{PL_i}{CCM} * 100\%, \quad PL_i \in CP, \quad (12)$$

where n is the number of posts, PL_i is the number of likes of *i*-post.

2. Talk Rate (*TR*) is an average number of comments in terms of audience size:

$$TR = \frac{1}{n} \sum_{i=1}^{n} \frac{PC_i}{CCM} * 100\%, \quad PC_i \in CP. \quad (13)$$

3. Amplification Rate (*AR*) is an indicator that determines the interest of users in the topic of a certain post:

$$AR = \frac{1}{n} \sum_{i=1}^{n} PR_i * 100\%, \quad PR_i \in CP. \quad (14)$$

4. Engagement Rate (*ER*) – the high level of user engagement indicates the quality and demand for the resource:

$$ER = \frac{1}{n}\sum_{i=1}^{n}\left(PL_i + PR_i + PC_i\right) * 100\%, \quad PL_i, PR_i, PC_i \in CP, \tag{15}$$

5. Engagement Rate by Reach (*ERR*) – the metric shows the ratio of users who have interacted with posts at least once to views:

$$ERR = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(PL_i + PR_i + PC_i\right)}{PV} * 100\%, \quad PL_i, PR_i, PC_i \in CP, \tag{16}$$

6. Engagement Rate of Post (*ER Post*) is the indicator that allows to evaluate the attractiveness of a certain post:

$$ER\ Post = \frac{PL + PR + PC}{CCM} * 100\%, \quad PL, PR, PC \in CP, \tag{17}$$

7. User Generated Content (*UGC*) – the metric allows to estimate the number of posts made by the users of the community [7]:

$$UGC = \frac{x}{n} * 100\%, \tag{18}$$

where *n* is the number of all posts, *x* is the number of posts, made by the users of the community.

Metrics for evaluating communication from SMM specialists

1. Post Rate is the number of posts in the community during a reporting period:

$$Post\ rate = \frac{N}{x} * 100\%, \tag{19}$$

where *n* is the number of posts for *x* days.

2. Response Time is a metric that shows the average administrator response time to user messages. This is an important indicator of the service quality and respect for cliens [8]:

$$Response\,time = \frac{1}{n}\sum_{i=1}^{n} x_i - y_i * 100\%, \tag{20}$$

3.3 Visual Analysis

There are various methods for finding hidden patterns using algorithms and machine learning, but do do not miss the opportunity to analyze and interpret data with the help of humans. Visual data analysis allows to represent large amounts of data graphically such as two-dimensional and three-dimensional graphs, tables, and decision trees.

This type of analysis has the following advantages:

- allows to analyze noisy data, as opposed to automatic methods that may not work well with such data;
- visual analysis does not require the implementation of complex algorithms;
- intuitive.

4 Research Results

Definition of text sentiment

During the study it was necessary to find out the most accurate method for determining the text sentiment. To train the models, RuTweetCorp data set was selected, which includes comments distributed into two groups: known positive (114.911 records) and known negative (111.923 records) [9,14].

The results of testing various models of text sentiment analysis are shown in Fig. 1.

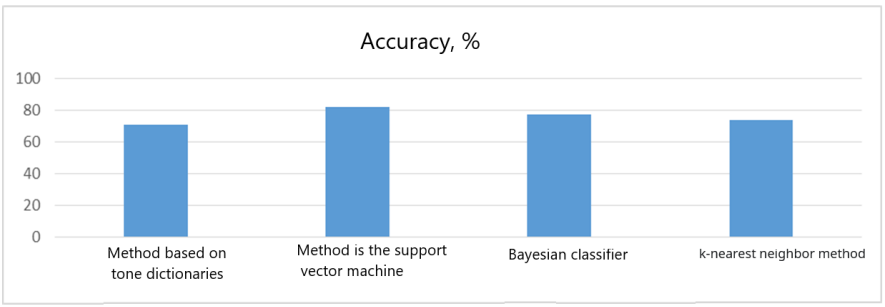

Fig. 1. Accuracy of models for definition of text sentiment

As Fig. 1 shows, the most accurate method is the support vector machine with 82%, the worst results were shown by the method based on tone dictionaries with accuracy of 71%. Thus, it is recommended to use the support vector machine to determine the text sentiment.

Definition of the target user profile

Fig. 2 shows the projection of user data (gender, age, country, city, interests, number of subscribers) into a two-dimensional space, and the clusters into which the data were divided using k-means cluster algorithm.

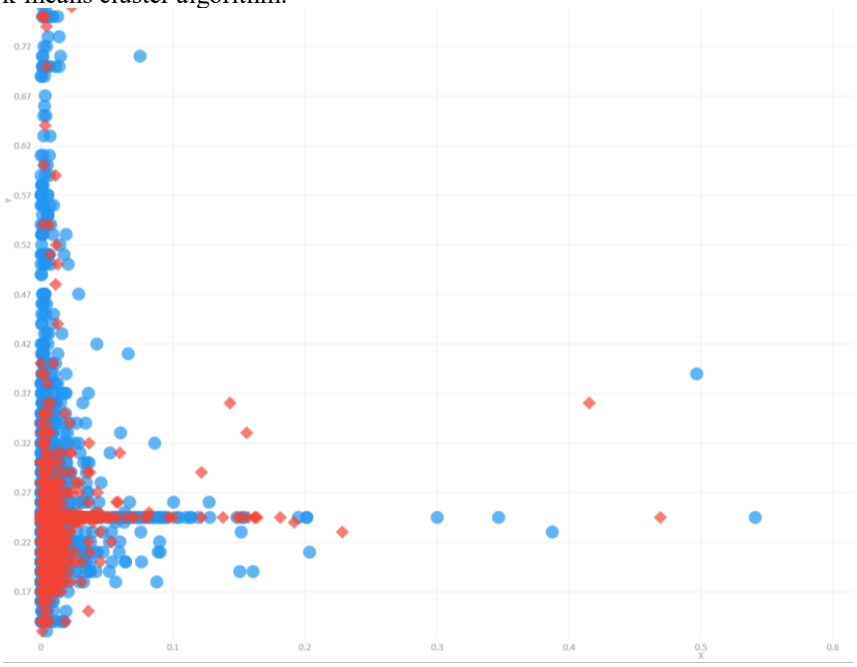

Fig. 2. Clustering community users in a two-dimensional space

Fig. 2 shows that users were divided into two clusters, and the data about each user were projected into a two-dimensional space. As you can see, clusters overlap – this is a normal phenomenon when analyzing users of the same community.

Fig. 3 demonstrates the distribution of clusters between the city of residence and the user's gender.

As Fig. 3 shows the condition for maximizing the distance to clusters is to divide the audience by gender.

Based on the information received, we can calculate the average values for each cluster, and determine the characteristics and size of the audience for the target advertising campaign.

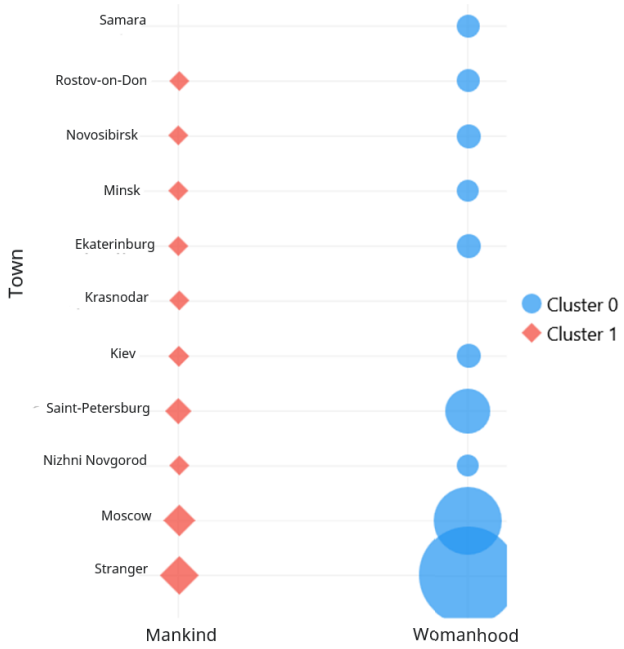

Fig. 3. Clustering community users according to gender and city of residence

Search for association rules

The search for association rules was performed using Apriori algorithm. Interests of users were used as input data of the algorithm. The search for association rules was carried out based on the interests of users of one of Vkontakte communities.

Most of the resulting rules with high support (C) and fidelity (S) are trivial, for example: Creativity → Humor (C = 0.94 and S = 0.65) or Photography → Humor (C = 0.92 and S = 0.27). However, the existence of these dependencies could be assumed without searching for association rules. Also there were found more interesting dependencies, for example: Online media, Education → Humor (C = 0.93 and S = 0.05), Humor, Photography, Literature → Creativity (C = 0.77 and S = 0.04).

On the basis of obtained data it is possible to search for a new audience in relevant communities, connected by interests.

Visual analysis

Visual representation allows to analyze large amounts of information quickly, for example, for one community of the social network Vkontakte, which is a mass media, a visual analysis was conducted (Fig. 4).

This example of visual analysis consists of six graphs and diagrams, which include: activities, countries and cities of residence, distribution by gender, distribution by gender and age, and a diagram of community user interests.

According to these charts, the following conclusions can be drawn: the majority of users live in Russia, more than 80% of the community's audience are men, the main age audience is from 18 to 21 years old, users who subscribe to this community are interested in humor, creativity, online media, education and video games, more than 84% are active users of social networks.

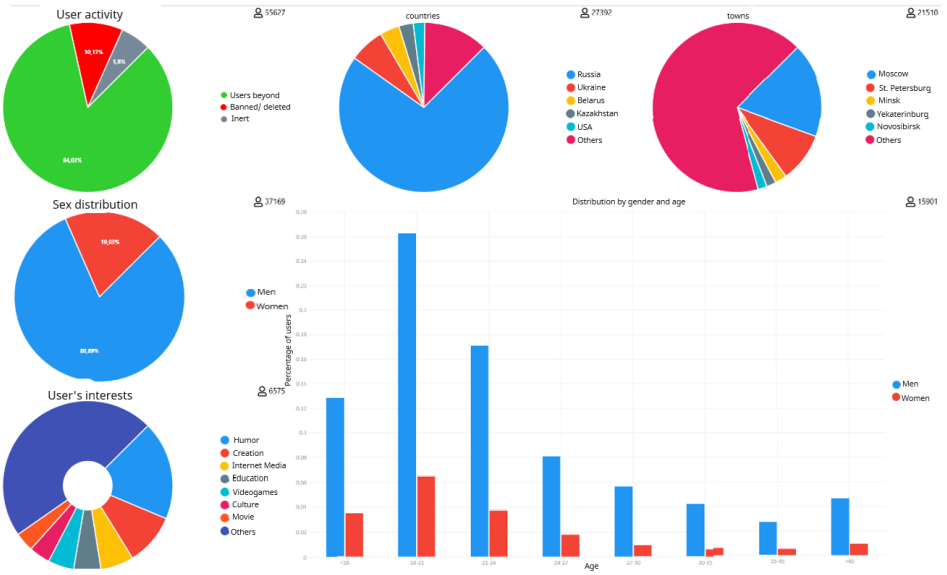

Fig. 4. Visual analysis of the community of social network Vkontakte

5 Conclusion

The result of the study is an information and analytical system for the analysis of social networks, which is designed to automate the collecting, monitoring and updating information on objects in social networks.

The components of this system are:

1. Data collection. The system can have many data sources, such as social networks, blogs, and websites.

2. Data transformation. At this stage, data from different sources are combined into a single storage system. Also, at this stage, the data are cleaned, scaled, and encoded.

3. Analytical processing. At this stage, data are automatically aggregated using OLAP technology [10]. The reason for using OLAP is the high speed of data processing. The structure of relational databases is convenient for operational databases (OLTP systems), but complex multi-table queries are performed relatively slowly.

4. Intellectual processing. Combines innovative approaches to data processing, such as Data Mining, Big Data, and Machine Learning.

At this stage, the sentiment of comments in the community and under posts is determined. This can be useful for research in the field of social media marketing.

Clustering community users will help to determine the target audience profile. Based on the results obtained, a corresponding target advertising company can be made, using data on the geographical location, age, and interests of the target user.

By searching for association rules, dependencies can be found in the interests of users.

On the basis of the statistical analysis, KPI indicators of communities were calculated, which determine the numerical indicators of effectiveness of community activities.

5. Presentation of the results. At this stage, the results found are presented in the form of graphs, diagrams, tables, and other visual objects. Visual analysis allows to transform complex data into visual images that enable the user to identify dependences.

One of the most promising areas of social media analytics is image analysis and semantic text analysis. The development of these areas will help better understand the interests and mood of the user and build more accurate analytical models.

Further development of the system can contribute to the creation of new models based on machine learning, allowing to make reliable predictions about different indicators, for example, determining the reaction of users to certain posts.

References

1. D.A. Gubanov, Social Networks. Models of information influence, management and confrontation. Moscow: Izdatelstvo phisiko-matematicheskoy literaturi, 228 p., 2010.
2. I.A. Chubukova, Data Mining. Moscow: Internet – Universitet Informatsionnich Tekhnologiy, 470 p., 2016.
3. E.A. Lebedeva, Analysis of message sentiment in microblogs using probabilistic models, 35 p.,2014.
4. E.V. Kotelnikov, Automatic analysis of text sentiment based on machine learning methods. Komputernaya lingvistika I intelektualtie tekhnologii. Materiali ezhegodnoi mezhdunarodnoi konferentsii Dialog [*Computational linguistics and intelligent technologies. Proceedings of annual international conference Dialogue*], **Issue 11 (18),** pp. 7–10., 2012.
5. Yu.A. Osipov, Application of k-means cluster analysis for classification of scientific texts. MSiM, 2017.
6. J. Hipp, U. Guntzer, G. Nakaeizadeh, Algorithms for Association Rule Mining [*A General Survey and Comparison. In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*], 2000.
7. R. Ramezani, M. Saraee, M.A. Nematbakhsh, *MRAR: Mining Multi-Relation Association Rules*, *Journal of Computing and Security*, **T. 1, № no. 2,** 2014.
8. I. Ivanichev, KPI in SMM. 30+ metrics of effective marketing in social networks. Internet – agentsvo Teksterra, 2018. Available at: https://texterra.ru/blog/kpi-v-smm-metriki-effektivnosti-marketinga-v-sotsialnykhsetyakh.html (accessed 25 October 2020).
9. A.A. Senatorov, Content Marketing: Social media promotion strategies. Moscow: Alpina Publisher, 160 p., 2020.
10. Yu. Rubtsova, Automatic construction and analysis of short texts corpus (microblog posts) for the task of developing and training a tone classifier. *Inzheneriya znaniy I tekhnologii semanticheskogo veb*a, **Vol. 1.** , pp. 109-116, 2012.
11. V.S. Belov, Information and analytical systems. Design and application basics. Moscow: Evraziiskiy otkritiy institut, 112 p, 2010.
12. E.A. Leonov, Y.A. Leonov, Y.M. Kazakov, L.B. Filippova, Intellectual subsystems for collecting information from the internet to create knowledge bases for self-learning systems.*Proceedings of the Second International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'17). IITI 2017*. Advances in Intelligent Systems and Computing. —2017— *vol 679*. — Springer, Cham, p. 95-103
13. A. A. Kuzmenko, D. E. Kondrashin, Methods and approaches to the development of a system for automated analysis of the dynamics of changes in the area of forest stands based on pattern automatic recognition methods. *ERGODESIGN*, *№ 4 (6*),230-240 pp, 2019
14. YU.A. Leonov, E.A. Leonov, A.A. Kuzmenko, A.A. Martynenko, E.E. Averchenkova, R.A. Filippov Selection of rational schemes automation based on working synthesis instruments for technological processes. Yelm, WA, USA: Science Book Publishing House LLC, 192 p., 2019.