Bull. Nov. Comp. Center, Comp. Science, 35 (2013), 1–22

© 2013 NCC Publisher

Methods for analysis of data from social networks

T. V. Batura, N. S. Kopylova, F. A. Murzin, A. V. Proskuryakov

Abstract. This work focuses on the analysis of online social networking services. We examine several formal definitions of various characteristics (numerical and structural) and introduce appropriate concepts, models and methods that could be useful for the analysis of information obtained from social networks. Preference analysis is proposed to be used to study interpersonal relations. Various modifications of Latané's dynamic social impact theory are proposed. The authors have studied techniques used to carry out psychological operations, and, as a result, have proposed several formal models. The paper briefly describes a software system that we have developed, which allows extracting, processing, analyzing and visualizing data from online social networking services. Our system has a data extraction module that can retrieve information from social networks such as Twitter, Facebook, and VKontakte.

Keywords: social network analysis, data mining, Latané's social impact theory, psychological operations.

1. Introduction

In recent years, online social networking services are becoming an increasingly popular means of communication. In addition to being convenient for communication, they open up new opportunities for the analysis of information flows and human behavior in the process of communication. It has been observed that the combination of the analysis of the social graphs structure with text mining in social networks provides the best results in studying the interactions between the members of the network [1, 2].

In this paper, we offer several numerical characteristics, graphs and sets that can be calculated or constructed on the basis of information obtained from social networks. The so-called preference analysis [3, 4] is proposed to be used for the analysis of interpersonal relations. The authors have made an attempt to adapt the dynamic social impact theory [5-7] to social networks. This approach could help to calculate the level of impact of other people on the opinion of a selected person.

The analysis of interactions between users requires the solution of some questions. How do messages affect the environment of their author? Which group of users is most affected by posts? For what purpose were they created? In order to answer these and other questions, we examined the methods of conducting psychological operations [7-10] and proposed a formal approach based on a modification of Latané's theory [11].

It is intuitively clear that relations between users and the vocabulary of text messages are interconnected. In the sixth section of the paper, we suggest a mathematical description of some sets and functions concerning the relations and vocabulary. The most important characteristics of these sets are considered.

A software package was created for extracting, processing and analyzing the user data. In particular, the system has a module which allows us to extract the data of the largest social networks (Twitter, Facebook, Vkontakte). This module is expandable to almost any social network, depending on the provided API.

Text normalization is performed using either the Porter stemmer algorithm [12], which has a high processing speed, or a morphological normalizer based on automatic text processing algorithms [13].

For data storage, the document-oriented database MongoDB [14] is used. For working with graphs, data can be imported into Gephi [15]. A data visualization module gives a possibility to plot the relationship between various parameters.

2. Numerical characteristics, relations and sets, computable on the basis of data received from social networks

When analyzing social networks, it is expedient to consider a set of numerical and non-numerical characteristics, relations and sets naturally connected with the users of a network and messages circulating in it. Importantly, all of them should be constructive, i.e. it should be possible to calculate or construct them by means of corresponding algorithms, naturally, in the presence of software allowing "importing" the necessary information from a network.

Let us designate by T the message ("twit") of a social network and by u, the user of a network who can create and send messages.

One-place characteristics:

Followers_Count (u) is the number of people who read messages of a given user (i.e. subscribers of this user);

$Friends\_Count\,(u)$ is the number of friends of a given user (the user himself adds some people to the list of friends);

Retweets (T) is the number of transfers (reposts) of a given message;

$Timeline\_Count\,(u)$ is the number of messages ("twits") created by a given user;

Real_name (u) is the "real" name of a user if the corresponding position is filled (it is a non-numerical characteristic).

Temporal characteristics:

Initiated (u) is the date of an account creation;

Datetime (T) is the date of a message creation.

Some sets:

$Followers\,(u)$ is a set of subscribers of a given user;

$Friends\,(u)$ is a set of friends of a given user;

$Mentions\,(u)$ is a set of names of users mentioned in the messages of a given user;

$Hashtags\,(u)$ is a set of hashtags present in the messages of a given user;

$Urls\,(u)$ is a set of external references (hyperlinks) present in the messages of a given user.

Numerical characteristics associated with sets:

$Count\_Mentions_u\,(v)$ is the number of mentions of the user v by the user u;

$Count\_Hashtags_u\,(v)$ is the number of mentions of the hashtag v by the user u;

$Count\_Urls_u\,(v)$ is the number of mentions of the external reference (hyperlink) v by the user u.

$Count\_Retweets_u\,(u_1)$ is the number of messages received from the user $u_1$and resent by the user u.

A hashtag is a word or phrase without spaces prefixed with the symbol "#". It is a form of metadata. Short messages in microblog systems and in social network services, such as Facebook, Twitter or Instagram, may be tagged by putting "#" before important words or phrases without spaces. They can appear in any sentence. For example, "New artists announced for #SXSW 2012 Music Festival!". Hashtags provide means of grouping such messages. One can search for a hashtag and get the set of messages that contain it.

Let us note that, in terms of the Latané social impact theory [5, 6], the function $Count\_Mentions_u\,(v)$ can be considered as "the power of influence" of the user v on the user u. The elements of the set $Mentions\,(u)$ can be ordered with respect to the characteristic $Count\_Mentions_u\,(v)$ in such a manner that the user mentioned most often will come first, the user mentioned by u a fewer number of times will be second, etc.

So $u_i \in Mentions\,(u)\,,\ i = 1, ..., N$ implies $Count\_Mentions_u\,(u_1) \geq$
$\geq Count\_Mentions_u\,(u_2) \geq ... \geq Count\_Mentions_u\,(u_N)$.

Usually, the analysis of processes occurring in groups of people involves three-place relations of preference [3]. The notation $i \vdash_{\overline{k}} j$ means, according to [4], that i is preferable to j in the opinion of k. Criteria of preference can be most various: professionalism in different kinds of activity generating a spectrum of new criteria; the ability to lead people; communication skills; susceptibility to innovations; psychological stability, etc. On the basis of these relations, an informal structure of the group is created, and in some cases, the importance of revealing this structure does not require any comments.

In paper [7], the campaign of poliomyelitis vaccination in India is described as based on messages (with the help of radio, print media, TV, cinema and posters) having influence on a society, and thereby TV and radio played the most important role in urbanized areas. In our case, it is possible to consider the power of the hashtag influence and the power of the reference influence as a remote analogue of mass media.

Three-dimensional relations:

$Mentions_u(u_1, u_2)$ means that the user u mentions the user u_1 not less often than u_2;

$Hashtags_u(h_1, h_2)$ means that the user u uses the hashtag h_1 not less often than the hashtag h_2;

$Urls_u(url_1, url_2)$ means that the user u mentions the reference url_1 not less often than url_2;

$Retweets_u(u_1, u_2)$ means that the user u resends the messages received from the user u_1 not less often than the messages received from the useru_2.

Numerical characteristics associated with three-dimensional relations:

$N\_Mentions_u(u_1, u_2) = Count\_Mentions_u(u_1) - Count\_Mentions_u(u_2);$
$N\_Hashtags_u(h_1, h_2) = Count\_Hashtags_u(h_1) - Count\_Hashtags_u(h_2);$
$N\_Urls_u(url_1, url_2) = Count\_Urls_u(url_1) - Count\_Urls_u(url_2);$
$N\_Retweets_u(u_1, u_2) = Count\_Retweets_u(u_1) - Count\_Retweets_u(u_2).$

It is possible to suppose that the functions mentioned above allow us to define the power of influence for various factors. For example, the function $N\_Mentions_u(u_1, u_2)$ allows us to find the power of influence of the user u_1 on the user u in comparison with u_2, i.e. the power of influence taking into account the preferences of the user u. For example, if the user u= '@navalny' mentions the user u_1= '@KSHN' 7 times and the user u_2= '@kudriavtsev' 1 time, then $N\_Mentions_u(u_1, u_2) = 7 - 1 = 6$. It is not clear yet how informative the function $Retweets_u(u_1, u_2)$ is.

3. The analysis of preference relations

In this section, some brief information from paper [4] is given with the purpose to show how the analysis of preferences is carried out. This problem is solved in two stages.

The first stage includes the information gathering. As a result, the individual questionnaires are formed. The questionnaire contains information on a pair ranging of the members of a group by the given criterion. More precisely, two members of the group are compared by a third member with whom the questionnaire is associated. From the mathematical point of view, a Boolean matrix corresponds to the individual questionnaire. As a result of considering the set of all questionnaires, we obtain the three-place relation.

The second stage includes information processing. An algorithm transforming a family of Boolean matrices (questionnaires) into a weighted graph is used. The structure of such graph reflects, to a certain extent, the structure of a group and can be analyzed by layers, depending on the weights of edges. It is supposed that the maximum positive response is guaranteed to the sender of an impulse when connected to the so-called “path of the greatest liking” or, by physical analogy, to “the breakdown path”. Such paths can be allocated in the graph mentioned above. Also important is the allocation of the first addressee of the impulse, i.e. the person through whom the group is “entered”.

In the process of the analysis of social networks, natural relations of preference appear. Namely, they are the three-place (three-dimensional) relations mentioned in the second section. It is interesting that, unlike sociologists, we compare not only people, but also hashtags and references.

3.1. The description of data and algorithm

Let us introduce the following designations: n is the number of members in a group; A_k is an individual questionnaire $(1 \leq k \leq n)$. The individual questionnaire is a Boolean (containing only zero and units) antisymmetric matrix with zeroes in the main diagonal, i.e. we have

$$A_k = (a_{ij}^k), (1 \leq i, j \leq n) a_{ij}^k = 0, (i \neq j \to a_{ij}^k = \bar{a}_{ji}^k),$$

where the upper line designates a negation $(\bar{0} = 1, \bar{1} = 0)$. The relation $i \underset{k}{\vdash} j$ corresponds to the value of an element $a_{ij}^k = 1$ of the matrix A_k. A collection of such questionnaires A_k, $(1 \leq k \leq n)$, forms the input data set.

An oriented graph G_k corresponds to each individual questionnaire A_k

$$G_k = \langle G_k, I_k \rangle, G_k = \{1, \ldots, n\}, \langle i, j \rangle \in I_k \leftrightarrow i \underset{k}{\vdash} j.$$

The matrix A_k is an adjacency matrix for G_k. The resultant matrix $Q = (q_{ki})$ is defined by the rule

$$q_{ki} = \sum_{j=1}^{n} a_{ij}^k,$$

i.e. the k-th line shows “the total opinion” of the k-th member about the i-th member of the group.

Further, $Opinion = \sum_{i=1}^{n} q_{ki}$ is the opinion of the k-th member about the group as a whole. The value $Rating = \sum_{k=1}^{n} q_{ki}$ is called the rating of the i- th member of the group and reflects the total opinion of all group members about the given one.

The graph $G = \langle G, I, w \rangle$, $G = \{1, ..., n\}$, $\langle i, j \rangle \in I \leftrightarrow i \neq j$ corresponds to the matrix. The edge weight is defined by the formula $w(i, j) = q_{ij}$. Let us note that the given graph is complete. Every two nodes are connected by a pair of edges oppositely oriented which can have different weights. Thus, the algorithm of the graph construction, which is our purpose, is completely described.

3.2. The analysis of the constructed graph

To analyze the constructed graph "by layers", we consider subgraphs containing the edges with weights larger than some threshold. Assume that for each natural number t we have

$$G^t = \left\langle G, I^t \right\rangle, I^t = \{\langle i,j\rangle \in I | w(i,j) = t\},\ \bar{G}_t = \bigcup_{s\geq t} G^s = \left\langle G, \bigcup_{s\geq t} I^s \right\rangle.$$

It is obvious that $t_1 \leq t_2 \to \bar{G}_{t_1} \supseteq \bar{G}_{t_2}$, $\bigcup\limits_{s\geq 0} G^s = G$.

The graph $\bar{G}_t$ is called a cut of the level t. Normally, formation of more powerful communications requires more time, though they can be formed very quickly sometimes, for example, as a result of the invitation of a highly skilled expert from outside. Therefore, considering G^t for various t, it is possible to see the dynamics of the communications development with time.

To find "paths of the greatest liking", it is necessary to solve some variant of a traveling salesman problem. It is not so convenient, if the graph contains a big number of nodes. In this case, it is expedient to do the following. At the beginning, we consider a cut at an appropriate t. Some nodes in it appear isolated and they are rejected. The obtained graph is supplied with edges to become a complete graph and weights are inherited, i.e. only significant members of the group are considered, but all communications between them are taken into account.

More precisely, we now study the graph $H^t = \{i \in G | \exists j (w(i,j) \geq t)\}$. Not to complicate designations, we suppose that we work with the initial graph.

Let $l = \langle i_1, \dots i_k \rangle$ be a route (a chain of edges) in a graph G. The weight of the route l is calculated by the formula

$$w(l) = \sum_{j=1}^{k-1} w(i_j, i_{j+1}).$$

Let us remind that the route is called closed if $i_1 = i_k$. We will use the following designations: $k \in l$ means that the node k belongs to the route l; $l_1 \subseteq l_2$ means that the route l_1 is a part of the route l_2; $Ent(k,l) = \{i | \langle k,i \rangle \subseteq l\}$ is an entry from the node k in the route l.

The traveling salesman problem (one of its variants) can be formulated as a problem of searching the maximum closed route without self-crossings, such that the value $w(l)$ reaches its maximum, i.e. the route should pass through all nodes. Thereby, it should pass through each node only once. Such a route exists, taking into account the completeness of the graph G. But in a general case, there can be several such routes. We will designate a set of all such routes by $L(G)$. By definition,

$$Ent(k, G) = \{Ent(k, l) | l \in L(G)\}$$

is the set of entries from the node k in the graph G.

The elements of $L(G)$ are called “paths of the greatest liking”, and $Ent(k, G)$ shows to whom the k-th member of a group can direct his impulses to “be connected” to these paths.

Let us note one more circumstance. In reality, paradoxical chains can be found in which the transitivity of preferences is broken. For example: “a is better than b”, “b is better than c” but “c is better than a”. It is possible to enter absolute and relative measures of transitivity.

The absolute measure is defined by the formula

$$f(k) = Card\{\langle i_1, i_2, i_3\rangle \, | i_1 \underset{k}{\mapsto} i_2, i_2 \underset{k}{\mapsto} i_3, i_1 \underset{k}{\mapsto} i_3\},$$

where $Card$ is the number of elements in a set.

The relative measure can be defined, for example, as follows: $F_k = f(k)/n^3$.

Usually a large-scale infringement of transitivity testifies to instability of a group. As a rule, a small-scale infringement is always present. Moreover, in a big group, insignificant communications are formed and broken up regularly. When passing to the corresponding level, it is possible to eliminate them and to carry out a qualitative analysis.

4. Latané’s theory of social impact and its modification

Now consider how to adapt the dynamic theory of social impact, proposed by Latané, to social networks.

Latané emphasized the importance of three attributes of the relationship between the recipient and source of information [5,6]: *Strength* is a credibility, or status of the agents involved; *Immediacy* is the physical or psychological distance between agents; *Number of sources* impacts the target audience.

According to the dynamic theory of social impact, the level of influence exerted on the individual may be represented by the following formula

$I_i = -S_i\beta - \sum_{j=1,j\neq i}^{N} \frac{S_jO_jO_i}{d_{ij}^{\alpha}}$, where

I_i is the amount of social pressure exerted upon the agent i;

O_i is the opinion of the agent i (±1) towards a proposition, “+1” represents support for a proposition, and “–1” represents opposition;

S_i is the power or influence of the agent i ($S_i \geq 0$);
β is an agent's resistance to changes ($\beta > 0$);
d_{ij} is the distance between agents i and j ($d_{ij} \geq 1$);
α is the distance decay exponent ($\alpha \geq 2$);
N is the total number of agents.

The value of β is usually taken to be 2 to conform with the value used in Latané's research [5,6]. Larger values of β mean that agents will require greater amounts of social pressure to change their opinion, lower values of β correspond to less social pressure. The value of α is also taken to be 2. Larger values of α mean that increase in the distance between the source and the target audience requires a significant increase in the social pressure.

Latané called the distance d_{ij} "immediacy", noted that it is an attribute of a pair of agents, and considered it as a measure of the ease of communication between two agents. This value may include the age, national, religious and other differences. The formula for calculating the value of d_{ij} may take into account the physical distance, e.g. the distance between the cities where the agents live. The ease of communication follows the inverse square law [7]. Generally speaking, various approaches to the analysis of social networks are possible, including non-use of the physical distance.

We propose a modification of the Latané's formula for the analysis of social networks:

$$I_u = -\beta \cdot \sum_{i=1}^{N} Count\_Mentions_u\left(u_i\right) - \sum_{i=1}^{N} \sum_{\substack{j=2 \\ i>j}}^{N} \frac{Mentions_u\left(u_i, u_j\right)}{\rho^{\alpha}\left(u_i, u_j\right)},$$

where $\rho\left(u_i, u_j\right)$ is the distance between the users u_i and u_j. This formula takes into account all usernames mentioned by the user u. We believe that all of them have an influence on him.

The next formula takes into account the influence of only the most and the least mentioned users:

$$I_u^1 = -\beta \cdot \max_{i=1}^{N} \left\{Count\_Mentions_u\left(u_i\right)\right\} - \max_{i=1}^{N} \max_{\substack{j=2 \\ i>j}}^{N} \left\{\frac{Mentions_u\left(u_i, u_j\right)}{\rho^{\alpha}\left(u_i, u_j\right)}\right\}.$$

We can also consider the influence of only the most mentioned user and the next mentioned one:

$$I_u^2 = -\beta \cdot \max_{i=1}^{N} \left\{Count\_Mentions_u\left(u_i\right)\right\} - \min_{i=1}^{N} \min_{\substack{j=2 \\ i>j}}^{N} \left\{\frac{Mentions_u\left(u_i, u_j\right)}{\rho^{\alpha}\left(u_i, u_j\right)}\right\}.$$

For example, it can be set as "follower – follower of a follower – follower of a follower of a follower, etc."

Since these calculations are made relative to the user's preference, it is appropriate to take advantage of the French railway metric:

$$\rho\left(u_i, u_j\right) = \begin{cases} \left\|u_i - u_j\right\|, & u_i - u = \lambda\left(u_j - u\right) \\ \left\|u_i - u\right\| + \left\|u_j - u\right\|, & u_i - u \neq \lambda\left(u_j - u\right) \end{cases},$$

where λ is a constant, u is a fixed point through which the path between u_i and u_j must pass. The simplest way is to count the number of edges. It is possible to assign a weight to each edge. In our opinion, the weight of the edge is not so important, because currently a user u may not be a follower of u_j and, as a result, the distance between them will be large, but the user u may become his follower at the next moment, and the distance will be reduced.

The formula reflects the influence of the user u_i on u as relative to other users $u_j \in Mentions\left(u\right)$; moreover, u_i is mentioned more often or as often as other users $u_j \in Mentions\left(u\right)$.

External global influences (such as mass media) can also be included in the model [7] by adding the following term "$-O_i O_M S_{Mi}$", where S_{Mi} represents the strength or influence media messages have on the agent i ($S_{Mi} > 0$); O_M is the opinion of the media (± 1). Given the influence of the mass media, Latané comes to the final formula [6,7]

$$I_i = -S_i\beta - O_i O_M S_{Mi} - \sum_{j=1, j\neq i}^{N} \frac{S_j O_j O_i}{d_{ij}^{\alpha}}.$$

Due to its omnipresent nature, the external influence is modelled as another agent outside the environment, but at a distance of 1 to every other agent. The value of S_{Mi} varies depending on the individual, as each agent feels a different level of pressure exerted on him by the mass media. This value is similar to the degree of credibility of the agent to the messages received from external sources.

We propose to consider a "power of hashtags" and "power of URLs" in social networks instead of mass media influence. Then we obtain the next formula

$$I_u = -\beta \cdot \sum_{i=1}^{N} Count\_Mentions_u\left(u_i\right) -$$

$$- \sum_{i=1}^{|Hashtags(u)|} \sum_{\substack{j=2 \\ i>j}}^{|Hashtags(u)|} Hashtags_u\left(h_i, h_j\right) -$$

$$-\sum_{i=1}^{N}\sum_{\substack{j=2\\ i>j}}^{N}\frac{Mentions_u\left(u_i,u_j\right)}{\rho^{\alpha}\left(u_i,u_j\right)},$$

where all hashtags and all usernames mentioned by the user u are taken into account.

Similarly, we obtain the formula

$$I_u = -\beta \cdot \sum_{i=1}^{N} Count\_Mentions_u\left(u_i\right) -$$

$$-\sum_{i=1}^{|Urls(u)|}\sum_{\substack{j=2\\ i>j}}^{|Urls(u)|} Urls_u\left(url_i, url_j\right) -$$

$$-\sum_{i=1}^{N}\sum_{\substack{j=2\\ i>j}}^{N}\frac{Mentions_u\left(u_i,u_j\right)}{\rho^{\alpha}\left(u_i,u_j\right)},$$

where all URLs and all usernames mentioned by the user u are taken into account.

5. Psychological operations

This section is based on the work devoted to military techniques in advertising [8]. It shows that conducting psychological operations and promoting “peaceful” advertising use similar tools and techniques. They have similar issues and problems, too. Psychological operations are characterized by a thorough technological preparation of the campaign, substantial research base for the study of the target audience to which the action is directed. The cycle of psychological operations consists of three components, also common for advertising: assessment, planning and implementation, which also include testing of messages and checking the results. Psychological operations consist of the following stages: intelligence gathering, analysis of the target audience, product development, selection of media, media production, and distribution.

The ultimate goal of the operation is behavioral changes in the target audience. Psychological operations are divided into three types: strategic, operational and tactical. Strategic operations are generally designed according to long-term aims in support of the general strategic planning, with measurable effects becoming visible in the indefinite future. Operational operations are directed at the regional target audience and planned to change the audience behavior faster than the strategic operations. Operational operations demonstrate strategic and tactical characteristics. Tactical operations are prepared and executed in objective areas in direct support of military tactical operations. Identification of ethnic peculiarities of the audience is a particularly important component, since the specialists conducting psychological operations are interested in finding vulnerabilities of the target audience.

Creation of messages can be divided [8, 9] into the following stages: integration, conceptualization and development. Integration is aimed to connect the analysis of the target audience with the appropriate type of media. It should answer the following series of questions: who is the target audience; where is the target audience; what is the aim of the message; when will this message have maximum effect; what is the purpose of the psychological operation.

The next section briefly describes a formalized model of psychological operations proposed by the authors of this paper. The detailed description of this model is given in [11].

5.1. Conceptualization

Conceptualization transforms the target analysis and media selection into a workable plan. The following general techniques [9] apply to any message: Attracting attention, Making a feeling of credibility, Making memories, Arousing emotions, Repeating the message.

These techniques can be modeled as some external (abstract) agents. We denote

$S_{ij} = f\left(B_{ij}\right)$, where

B_{ij} is the amount of influence j on the agent i;

S_{ij} is efficiency of influence j on the agent i.

In the simplest case, we can assume that the function f is linear. However, the function reflecting that the audience rejects excessive information is closer to reality. Examples of the function f are shown in Figure 1.

Similarly to Latané's theory of social impact, the degree of influence on the agent i can be calculated by the formula

$-\sum_{k=1}^{L} \hat{S}_{ik}\beta_{ik}$, where

k is the influence external to the set of actions carried out in the course of psychological operations;

L is the number of all external influences ($k \leq L$);

$\hat{S}_{ik}$ is the external influence k on the agent i;

β_{ik} is the resistance of the agent i to the external influence k.

Psychological operations, as opposed to advertising, are characterized by interpersonal communication. Direct communication between people has many advantages: adjustment to the audience may be realized, frequent

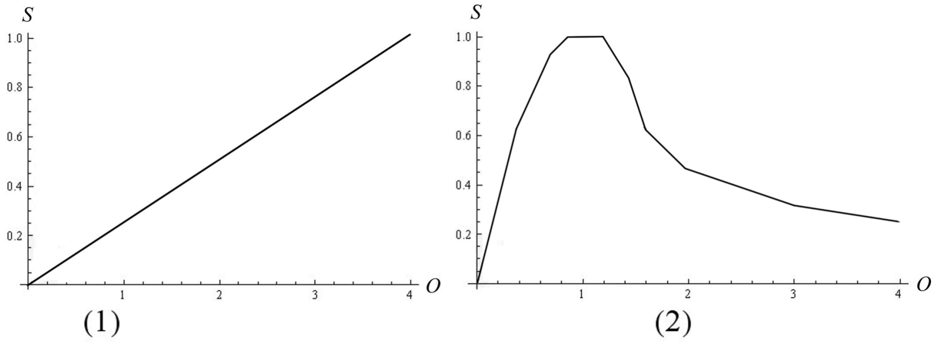

Figure 1. Examples of the function f

repetitions are allowed, the easier way of selection of a target audience is available, the methods of impact on the audience can sometimes be more effective than some media manipulations, etc.

The use of cellular automata for the modeling of social influence processes [7] requires the following: the entire audience is divided into subgroups each of which is considered as an agent. Social pressure exerted on the target audience is estimated on the basis of social pressure exerted on each agent and the degree of adjustment of the information to each subgroup. Clearly, all these quantitative characteristics and the form of the function f are defined by an expert on the basis of his personal experience.

5.2. Dealing with Rumors

One of the most important components of interpersonal communication is rumors. They are also used in advertising and psychological operations. Rumors are a powerful weapon, but they should be well-prepared and kept under control.

We can distinguish three characteristics of rumors.

1. The source must be appealing and credible to the audience. Suppose that an expert has determined the degree of rumor appealing $O_i^{(1)}$ for the agent i. In the simplest case $O_i^{(1)} = \pm 1$, where $O_i^{(1)} = 1$ means that a rumor is attractive, and otherwise $O_i^{(1)} = -1$. In the continuous case $O_i^{(1)} \in [0, 1]$.

2. The rumor content must be credible. Denote by $O_i^{(2)}$ the degree of confidence of the agent i. In the simplest case $O_i^{(2)} = \pm 1$, where $O_i^{(2)} = 1$ means that the agent i believes the rumor and otherwise $O_i^{(2)} = -1$. In the continuous case $O_i^{(2)} \in [0, 1]$.

3. The receiver of a rumor becomes a repeater when he transmits the rumor. Let $O_i^{(3)}$ be the activity of the agent i during the message transmission. Similarly to neural networks, the value S_{ij} is the activity of message transmission from the agent i to the agent j. The impact of the rumor on the agent i can be calculated by the formula $\alpha_i = \sum_{j \in J_i} S_{ij} O_i^{(1)} O_j^{(2)}$, where J_i is the set of agents having an influence on the agent i.

If α_i is greater than a certain threshold, then the agent i will change his opinion under the influence of rumors. Even if the agent does not change his opinion in favor of the rumor, he may spread it. The threshold β is an indicator whether the agent will transmit the message or not. The value of β is chosen empirically and depends on the audience. If $O_i^{(3)} \alpha_i > \beta$, the rumor passed on.

The methods of creating rumors are of special interest. The features of information transfer and characteristics of human perception are important for understanding how rumors are created and spread. There are three transmission characteristics in rumors spreading: leveling, sharpening, and assimilation.

Leveling refers to the loss of details during the transmission process. The focus is on the features of information processing by different social groups. One of the rules of this stage is as follows: information that is not leveled by some groups becomes important. Information sharpened by some groups may be leveled by others. The degree of leveling can be considered as a monotonically decreasing function of the length of the rumor (in words), given by an expert.

Sharpening (selective perception or perceptual selectivity) is selection of a limited number of details from the original text. A long message is broken into fragments, each of which is characterized by the degree of appealing $O_i^{(1)}$ for the agent i. Sharpening can be interpreted as follows: the fragment with a maximum of $O_i^{(1)}$ is selected and further transmitted. As a rule, the value of the confidence degree $O_i^{(2)}$ of the agent i decreases.

Assimilation reflects the existing stereotypes, prejudice and ethnocentrism, i.e. the subconscious motivations. The entire set of agents is divided into disjoint subsets

$\bigcup_{i=1}^{K} \Omega_i = \Omega$, where

Ω is the set of agents, and $i \neq j \to \Omega_i \bigcap \Omega_j = \emptyset$.

For example, we divide the set of agents according to ethnicity. The agents belonging to different nationalities have different emotional reactions to a particular rumor, and therefore different degrees of confidence and so on.

Now consider the behavior of the values S_{ij}, B_{ij}, and O_{ij}. The classification of types of crowds [10] is useful for this purpose.

A casual crowd is a loose collection of people with no real interaction (e.g, a crowd of spectators or people in the supermarket). It emerges spontaneously. The value S_{ij} is small.

A conventional crowd is a deliberately planned meeting (e.g., the audience of sporting events, concerts, and other performances). The value B_{ij} is small but S_{ij} increases.

An expressive crowd occurs when many people gather with a specific purpose of experiencing strong emotions (e.g., religious revival shows, celebrity funerals, etc). The values S_{ij} and O_{ij} are rather large.

An acting crowd is a crowd that wants to bring changes (e.g., political demonstrations, boycotts, strikes, and pickets). The value B_{ij} is small, but the values S_{ij} and O_{ij} are large.

5.3. The SCAME approach

SCAME (Source, Content, Audience, Media, Effects) is a convenient and very efficient method for the analysis of propaganda and counterpropaganda. The analysis of propaganda using the SCAME approach determines the source of the propaganda, the message content, the whole target audience, and the specifics of the medium used to send the message [9].

The source analysis examines not only propaganda but also the organization responsible for its development and dissemination (an individual, organization, or the government). The source analysis can help to determine the accuracy, credibility, kind of connections, etc. So we can define several sources $Source_1, ..., Source_k$ which play a special role of the source agents in our model. Besides, we have several expert agents $Expert_1, ..., Expert_s$. According to the opinion of $Expert_i$, each source is characterized by a vector $M_{ij} = \left\langle m_1^{(i,j)}, ..., m_t^{(i,j)} \right\rangle$, where $m_k^{(i,j)}$ are characteristics such as credibility, accuracy, and connection to the government, military command, organization, or individual. We can assume that $0 \leq m_k^{(i,j)} \leq 1$. The vector M_{ij} is called the independent opinion of $Expert_i$ about $Source_j$.

The content analysis evaluates what the propaganda message says. It also determines the motive and goals of the source. The content analysis reveals the meaning and the reasons of the message dissemination and the intended purposes of the message. We have different types of information: involuntary information, biographic information, economic data, propaganda inconsistencies, geographic information, and intentions. Let V_t be the amount of information of each type. If the information comes from several sources, it is possible to take a linear combination of them with weights (for each class separately). Moreover, it is possible to take into account the reliability of the source and the amount of information of each type.

The audience analysis involves the study of the whole target audience. It determines the reasons a particular audience was selected and the rationale for a particular line of persuasion. The audience can also be classified with respect to income, nationality, geography, ethnicity, political preferences, religion, race, social class, caste, and other factors. The audience analysis identifies three types of the target audience: *groups, categories*, and *aggregates. Groups* are collections of people bound together by common activities and goals. The groups are divided into primary and secondary. An example of a primary group is a family or a small military unit such as a squad or platoon. A primary group is extremely protective of its members from outside interference. An example of a secondary group is a parliament, united in its goals of serving the electorate and country but perhaps divergent in individual views for accomplishing its mission. *Categories* are collections of people who share specific demographics such as the race, sex, or age. *Aggregates* are collections of people identified by a common geographic area. Examples of aggregates are Europeans, Asians, Egyptians, etc.

According to another classification, there are four major types of audience: apparent, ultimate, intermediate, and unintended. The detailed description of the target audience classification can be found in [9]. A category can be considered as one integral agent. In this case, the size of each category is a characteristic of the agent. Combining the categories and considering them as a single agent, we significantly simplify modeling but, of course, this leads to a loss of accuracy.

Media analysis determines why a dissemination method was chosen, what media capabilities the opponent has, and how consistent the message content was. Messages could be received via audio, visual, and audiovisual means (newspapers, magazines, leaflets, radio, television, etc.), each of which is characterized by its effectiveness S_{ij} (the power of influence).

The effect analysis determines the overall results of the opponent propaganda. For instance, the results may include the specific effects of the propaganda on the target audience and the reasons it was effective, partially effective, or totally ineffective. Let *Effect* (*Source*, *Content*, *Audience*, $Media$) be the target function for a specific source, audience, content, and media. This function is constructed on the basis of an expert estimation and local testing. At the first step, the number of agents with a positive (+1) and negative (-1) opinion is found before the message appeared and after that. At the second step, we obtain $N = N^+ + N^-$, i.e. the total number of agents before the message appeared; M^+ and M^- are the number of agents with a positive and negative opinion after the message appeared. The value of $Effect = (M^+ - N^+)/N$ shows how the number of agents with a positive opinion increased. The following formula takes into account all conditions affecting the propaganda:

$\sum_{\{S,C,A,M\}} \alpha_{\{S,C,A,M\}} Effect\,(S, C, A, M)$, where

$\alpha_{\{S,C,A,M\}}$ is a value determined by the number and credibility of sources, the content, audience, media, and possibly the cost of media.

Let us say a few words about counterpropaganda. The counteraction techniques are used to reduce or deny the opponent's propaganda message. Each of these techniques has its positive and negative effects. These techniques include a direct and indirect refutation, diversion, silence, forestalling, and some others.

The direct refutation should be circulated as widely and quickly as possible providing the true information to the target audience before the original message has inflicted any lasting damage. This aggressive technique attracts the audience's attention and may give additional credibility to the opponent messages repeating them. So, this technique is suitable when the condition

$Effect\,(Source_{opp}, Content_{opp}, Audience, Media_{opp}) <<$

$<< Effect\,(Source_{contr}, Content_{contr}, Audience, Media_{contr})$ is satisfied.

The indirect refutation challenges the credibility of the opponent propaganda. The advantage of this technique is that it does not reinforce or spread the opponent propaganda as readily as the direct refutation. An example of this method is an independent campaign aimed at reducing the degree of confidence $O_i^{(2)}$of as many agents as possible.

A diversion tries to overshadow the content of the opponent message by presenting a topic that diverts the audience's attention from this message.

Silence. Sometimes it is better to remain silent than distribute a message from somebody else. There is a threshold time t_0, that indicates how long the agents remember the content of the message. If new messages are not obtained after some time $t \geq t_0$, the interest fades, the opinion becomes neutral, and the messages will not spread. Here it is important to evaluate a short-term effect of this message and understand how dangerous it is.

Forestalling anticipates the potential opponent propaganda. Using this technique, we counteract the subjects potentially exploitable by the opponent propaganda before the opponent seizes them for his own purposes, i.e. the value N^+ should be maximal at the beginning of the opponent's campaign.

6. Some interesting sets of users of a social network and their characteristics

In the analysis of social networks, the following situation is of interest for us: two graphs are selected according to some criteria. For example, suppose that there are two specialized dictionaries (thesauri) corresponding to certain fields of knowledge. Accordingly, there are two sets of users selected with respect to the type of lexicon in their messages. Let us note that the intersection of these sets can be nonempty even if the dictionaries do not intersect it consists of active users of both thesauri. Thus, the nodes of graphs are the sets of users. The edges correspond to mutual citing of users, or various criteria connected with their lexicon. The weight of nodes and edges may be defined by means of various frequency characteristics related to citing or lexicon. It is also interesting to consider not only the intersection of the given graphs or their symmetric difference, but also some other subgraphs, or more precisely, some subsets of nodes, for example, a set of nodes "close enough" to the intersection. The exact definitions are given below.

Assume we have two graphs $G_i = (V_i, E_i)$, $i = 1$, 2, where V_i is a set of nodes and E_i is a set of edges of a graph G_i. The sum $V_1 \bigcup V_2$, the intersection $V_1 \bigcap V_2$ and the symmetric difference $V_1 \Delta V_2 = (V_1 \bigcup V_2) \backslash (V_1 \bigcap V_2) = (V_1 \backslash V_2) \bigcup (V_2 \backslash V_1)$ are defined naturally.

Let us denote for brevity $H = V_1 \bigcap V_2$ and $adj(x, y) \leftrightarrow E(x, y) \vee E(y, x)$. Then $Adj(x) = \{y : adj(x, y)\}$ is a set of nodes adjacent to the node x; $AdjH(x) = Adj(x) \bigcap H$ is a set of nodes adjacent to x in the intersection H; $CAdj(x) = Adj(x) \backslash AdjH(x)$ is a set of nodes adjacent to x which are not in the intersection H.

Further we suppose that

$\omega_i : V_i \to N$ is the function setting the node weights;

$r_i : E_i \to N$ is the function setting the weights of edges.

It is possible to define the weight function on the sum of the graphs

$$\omega(x) = \begin{cases} \omega_1(x), \ if \ x \in V_1 \backslash V_2, \\ \omega_2(x), \ if \ x \in V_2 \backslash V_1, \\ \frac{\omega_1(x) + \omega_2(x)}{2}, \ if \ x \in V_1 \bigcap V_2. \end{cases}$$

The function $r : E_1 \bigcup E_2 \to N$ can be defined similarly.

Here are the numerical characteristics of some subgraphs:

$$\alpha_i = \sum_{x \in V_i} \omega_i(x), \lambda_i = \sum_{x \in V_1 \bigcap V_2} \omega_i(x), \lambda = \sum_{x \in V_1 \bigcap V_2} \omega(x), \mu = \sum_{x \in V_1 \Delta V_2} \omega(x),$$

$$\beta_i = \sum_{e \in E_i} r_i(e), \beta(x) = \sum_{y \in Adj(x)} r(x, y), \gamma(x) = \sum_{y \in CAdj(x)} r(x, y).$$

Let us consider the sets $V_i' = \{x' \in V_i : \exists x \in H \ (adj(x, x'))\}$. Of greatest interest is the set $L = (V_1' \bigcup V_2') \backslash H$ of those users of a network who are very close to the intersection of two communities of users.

The most interesting numerical characteristics are the following.

1. The *tolerance* of a node

$$T(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \sum_{y \in AdjH(x)} \omega(y).$$

2. The *degree* of protection of a node

$$D(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \sum_{y \in CAdj(x)} \omega(y).$$

3. *Compatibility* of the sets V_1 and V_2

$$Q = \sum_{x \in H} \omega(x) + \sum_{x \in V_1 \Delta V_2} T(x) - \sum_{x \in V_1 \Delta V_2} D(x).$$

The formulas considered above can be generalized in order to take into account the weights of edges.

1. The tolerance of a node taking into account the weights of edges

$$T'(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \frac{1}{\beta(x)} \cdot \sum_{y \in AdjH(x)} \omega(y) \cdot r(x, y).$$

2. The degree of protection of a node taking into account the weights of edges

$$D'(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \frac{1}{\gamma(x)} \cdot \sum_{y \in CAdj(x)} \omega(y) \cdot r(x, y).$$

3. Compatibility of setsV_1 and V_2 taking into account the weights of edges

$$Q' = \sum_{x \in H} \omega(x) + \sum_{x \in G_1 \Delta G_2} T'(x) - \sum_{x \in G_1 \Delta G_2} D'(x).$$

7. Program implementation

As a result of our research, a program complex has been developed that contains the modules of information extraction from social networks and its processing, analysis and visualization. All modules are implemented in the Python language for different operational systems on which the complex can work. The structure of the program complex is shown in Figure 2.

The data extraction module can take data, first of all, from the largest social networks: Twitter, Facebook, VKontakte. To access each of them, the applied programming interface (API) is used; authorization is made by means of the OAuth protocol.

The basic difficulties at the stage of data extraction are restrictions on the number of requests to servers of social networks from a certain IP address. At the initial stage of implementation, a restriction on access to Twitter servers was no more than 350 requests during one hour, and at present it is from 60 to 720, depending on the type of a request. The VKontakte servers have a restriction of three requests in a second. The Facebook servers have no restrictions on the number of requests, however each request is processed for 1-2 seconds. To overcome these difficulties, it was decided to use proxy servers (now it is possible to load their list into our module).

The user data received can be divided into three categories: 1) the user's personal data, such as the name, nickname, registration time; 2) user's messages; 3) communications between users. After information extraction, the data processing module carries out the search for markers, hashtags, mentions of users, references, etc. Further, normalization of the texts of messages is made depending on the options: with the help of the Porter's stemmer [12] (for higher processing speed), or the morphological normalization on the basis of the AOT algorithms [13] with the help of the PyMorphy library.

For data storage, the document-oriented database MongoDb [14] is used. The choice fell on this DB for the following reasons.

1. There are the mechanisms of data processing Map Reduce and Aggregation framework considerably accelerating the data processing in the DB.
2. The speed of the DB, namely, its writing time is 2.5-3 times less than that of MySQL, and its reading time is 1.7-2.5 times less, depending on the number of records.
3. The absence of any rigid structure of data (more exactly, variability of the structure) and the simplicity of changing it were an important condition, because at the initial stage there appeared a large number of unstructured data, and the structure of data storage was gradually developed in the process of their extraction and analysis.

In the module of the data analysis, various algorithms are used for clusterization and classification of users and their communications and messages. The module of a graph structure construction has a functionality for construction of graphs reflecting the users communications. Here different data can be used, both initial and received by the analysis. There is also an opportunity to upload data in Gephi [15], the software allowing effective work with graphs, both in a special format and by means of the HTTP protocol. The module of data visualization provides us with a tool for building the di-

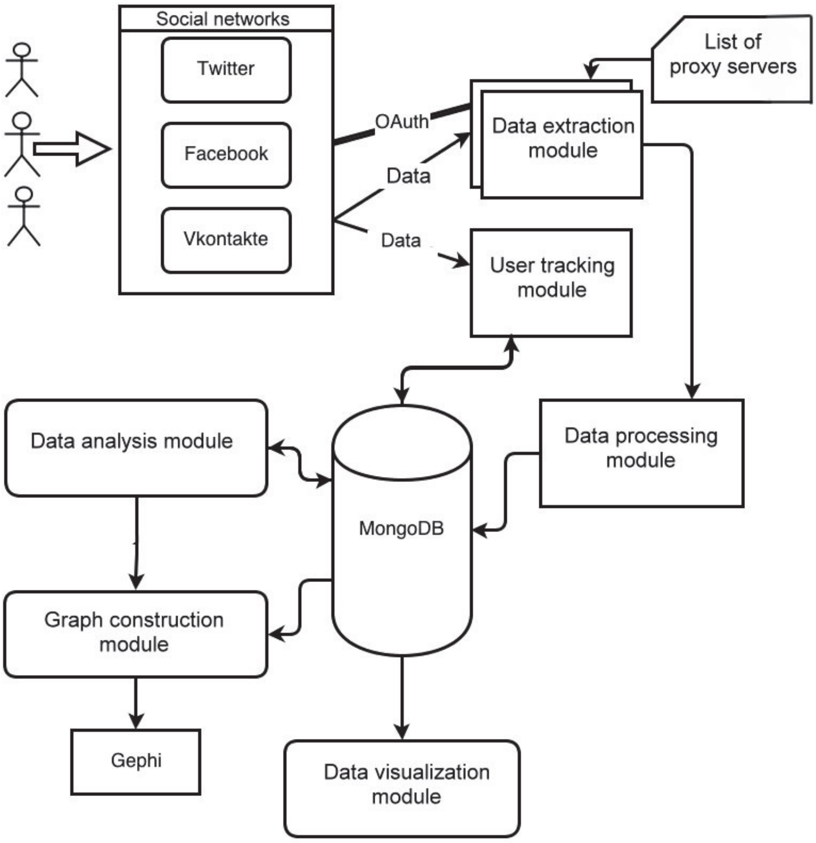

Figure 2. The structure of the program complex

agrams of dependences between various indicators on the basis of extracted data.

8. Conclusion

This paper is devoted to the problem of the analysis of online social networking services. We offered several numerical and structural characteristics, introduced some concepts, models and methods that could be useful for the analysis of information obtained from social networks. A software system has been developed that allows extraction, processing, analysis and visualization of data from online social networking services. Our system has a data extraction module that can retrieve information from social networks such as Twitter, Facebook, VKontakte.

The software package allows performing from 8 000 to 250 000 requests per day, using a single computer. The number of requests depends on the social network, the speed of hardware and bandwidth. Obviously, the amount of information received is very large and it will increase even more with an increase in the number of computers used (i.e., within a distributed system of data acquisition and processing). Therefore, first of all we have to select the part of information that can be processed effectively and needed for specific purposes. The easiest way is to use keywords. This method is implemented in our system. Actually, a more detailed study of other non-trivial methods is required.

If the volume of information is not so large, all features, concepts, and techniques described in this paper are highly constructive except those from Section 5. We mean that the numeric and nonnumeric characteristics, relations, sets, and graphs naturally associated with the network users and their messages can be easily calculated or constructed using appropriate algorithms.

Section 5 briefly describes a formalized model of psychological operations. The model is constructed on the basis of the previously published materials on this subject. We used a modification of Latané's theory of social impact and a multiagent approach. The proposed approach can be used in practice, but the experts will have a heavy workload (on specification of constants, functions, etc.). In this approach, the most time-consuming parts of testing are the selection of target groups for formation of a data sample and the development of a set of reasonable rules of conduct and impact on agents. If the data set and the rules of conduct are well established, the simulation will be more robust, realistic and credible but, of course, great efforts should be made to collect data.

References

[1] Social Network Data Analytics / Ed. by Ch.C. Aggarwal. – Springer, 2011.

[2] Batura T.V. Methods of Social Networks Analysis // Vestnik of Novosibirsk State Univ. Ser.: Information Technologies. – Novosibirsk. – 2012. –Vol. 10, Iss. 4. – P. 13–28 (In Russian).

[3] Rogers E.M., Agarwala-Rogers R. Communication in Organizations. – M.: Ekonomika, 1980 (In Russian, translated from: New York: Free Press, 1976).

[4] Kryuchkov V.N., Murzin F.A., Nartov B.K. An investigation of the connections in the collectives and in the computer networks // The Problems of Constructing the Efficient and Reliable Programs. – Novosibirsk. – 1995. – P. 136–141 (In Russian).

[5] Latané B. The psychology of social impact // American Psychologist. – 1981. – Vol. 36. – P. 343–356.

[6] Nowak A., Szamrej J., Latané B. From private attitude to public opinion: a dynamic theory of social impact // Psychological Review, 97. – 1990. – P. 362–376.

[7] Wragg T. Modeling the effects of information campaigns using agent-based simulation. – 2006. – 61 p. – (Prep. / Command and Control Division, Defense Science and Technology Organization, Australian Government; DSTO-TR-1853).

[8] Pocheptsov G. Military methods of the peaceful advertisement // Reklamnoye Izmereniye. – 1998. – Iss. 7(48) (In Russian).

[9] Allport G.W., Postman L. The Psychology of Rumor. – Oxford: Henry Holt, 1947.

[10] Blumer H. Collective Behavior // Principles of Sociology / Ed. by A. M. Lee. – N. Y.: Barnes and Noble, 1951. – P. 67–121.

[11] Kopylova N.S., Murzin F.A. Modelling of mechanisms of social influence on a basis of multi-agent approach // Questions of Artificial Intelligence (Vestnik of Scientific Council on Methodology of Artificial Intelligence of RAS). – 2009. – P. 173–183 (In Russian).

[12] Willett P. The Porter stemming algorithm: then and now // Program: Electronic Library and Information Systems. – 2006. – Vol. 40(3). – P. 219–223.

[13] Automatic Text Processing [WEB Resourse]. 2012.
Access Regime: http://aot.ru/ (In Russian).

[14] MongoDb [WEB Resourse]. 2012.
Access Regime: http://docs.mongodb.org/manual/reference/replica-status/ .

[15] Gephi, an open source graph visualization and manipulation software [WEB Resourse]. 2012. Access Regime: https://gephi.org/ .