Modeling of online social networks for automated monitoring system

Yu.B. Savva[a], Yu.V. Davydova[a]

[a] *Orel State University, 302026, 95 Komsomol'skaya street, Orel, Russia*

Abstract

Monitoring using keywords is necessary step in solving the problem of detection of users' illegal behavior in online social networks such as drug use propaganda. Analysis of text messages is difficult because of using jargon and making mistakes in communications. In this paper model of online social networks for automated monitoring system is presented. The main feature of this model is emphasis not on communications between users but on text messages.

Keywords: online social networks; monitoring; text analysis; information retrieval; fuzzy search

1. Introduction

Advantages of online social networks (OSNs) such as high speed of information dissemination which can be compared with a virus and ease of use make them a convenient tool for information influence and propaganda of deviant and illegal actions. Threats of OSNs, such as extremist and terrorist groups, were discussed in [1]. For providing information and psychological security of users, automated monitoring system of OSNs is required. Most existing monitoring systems [2] are used for business goals and find out users' attitude to brands. Sharing their opinions about products, service or social events users try to use hashtags – correctly written mentions with special label or metadata. It makes it easier for users to find messages with a specific theme or content. As for illegal activity, it isn't advertised or advertised for closed groups of users though propaganda can be an exception. It is more difficult to find messages connected with searching topic without hashtags especially if they are written with mistakes. Orthographic and typographic mistakes are common for informal writing on the whole and for text messages in OSNs in particular. Also informal writing is often characterized by using slang and different abbreviations. As for illicit fields of activity, communications there often contain specialized jargon. We consider jargon as a highly specialized slang, which is often used in closed communities and is hard to understand. These and some other features of text messages in OSNs were discussed in [3], taking them into consideration it can be said without doubt that monitoring process is difficult and requires special methods for analysis of text messages.

Process of monitoring involves a kind of information retrieval, text messages from OSNs are gathered and fuzzy search by keywords is organized. Keywords represent lexics, which is used in communications in illegal fields of activity. This paper describes model of online social networks used in automated monitoring system. According to system's goal, the emphasis of the model is made on text messages.

2. The object of the study

Usually online social network is defined as a graph $G(N,E)$, where $N=\{1,2,..n\}$ is a set of vertices (agents - users, communities) and E is a set of edges which represents interaction of agents [4]. Tasks of users' behavior modeling, users' interaction modeling, analyzing features of subgraphs of friendship are popular. The main goal of current automated monitoring system is decision support in detection of illegal behavior in OSNs, wherein information retrieval and analysis of text messages play a big role. Thus, text messages should be included in model.

Let us denote $I=\{i_1,i_2,...,i_{ic}\}$ is a set of identifiers of OSNs users or communities, where ic is the number of identifiers. $M=\{m_1,m_2,...,m_{mc}\}$ is a set of messages, mc is the number of messages. Messages are gathered into groups: $M=\bigcup_{k=1}^{ic} M_k$. Every message can be represented as follows:

$$m_j=\langle i_k, text_j, t_j, type_h, parent_j\rangle,\ i_k\in I,\ j=\overline{1..mc},\ h=\overline{1..3},$$

$$parent_j=\begin{cases}\varnothing, type_j=type_1\\ i_n\in I, n=\overline{1..ic}\end{cases},$$

where: – i_k is identifier of user who posted the message m_j;

– $text_j$ is text of the message m_j, $text_j = \langle w_{j1}, w_{j2}, ..., w_{jg} \rangle$, w_{ji} is the *i*-th word in text;

– t_j – date and time of the posted message m_j;

– $type_h \in Type$, $Type = \{type_1, type_2, type_3\}$ is a set of message types, where $type_1$ is original message (which means that user who posted message is its author), $type_2$ is reposted message (which means that user posted somebody's message), $type_3$ is a comment to original or reposted message;

– $parent_j$ is a user's or community's identifier. If current message is reposted message or comment then *parent* contains identifier of author who posted original message.

$P = \{p_1, p_2, ..., p_{ic}\}$ is a set of pages of OSNs, number of pages is equal to number of users' and communities' identifiers as every page belongs to user or community. Page is defined as follows:

$$p_k = \left\langle i_k, tt_q, c_k, M_k = \left\{ m_{kz} \middle| z = \overline{1..x} \right\} \right\rangle, i_k \in I,\ x < mc,\ q = \overline{1..2},$$

$$c_k = \begin{cases} \varnothing, tt_k = tt_1 \\ \{i_n\} \subset I, n = \overline{1..ic} \end{cases},$$

where: – i_k is the identifier of user or community of current page p_k;

– $tt_k \in TT$, $TT = \{tt_1, tt_2\}$ is a set of pages type. tt_1 is a personal page and tt_2 is a community page;

– c_k is a set of user's identifiers. If current page p_k is a personal page then c_k is an empty set as page p_k belongs to one user. If p_k is a community page then c_k keeps user's identifiers who are owners or managers of community (it can be one user, so c_k keeps one element);

– M_k is a group of messages which are posted on the page p_k. It can be empty $M_k = \varnothing$, that means OSNs page doesn't contain any messages at the moment.

Set of keywords is given $L = \{l_1, l_2, ..., l_{lc}\}$, where *lc* – the number keywords. Every keyword is represented by its grammatical, semantic information and word forms (according to inflection rules in Russian language) $l_s = \{GR_s, SM_s, WF_s\}$. This keywords storage model was described in [5]. In this work we are focused on word forms of keywords. They was defined as a language *WF* over the alphabet *A*, $A = \{à, á, â, ..., ÿ\}$. $WF \in A^+$.

The goal of automated monitoring system of OSNs is to find set of pages $PF \subset P$ which contains required amount of keywords, therefore these pages are indicators of potential illegal actions of their owners. Conceptually it can be presented as follows: $PF = \left\{ \left\langle i_k, tt_q, c_k, M_k = \left\{ m_{kz} \middle| z = \overline{1..x} \right\} \right\rangle \middle| \left(\sum_{z=1}^{x} \sum_{q=1}^{lc} f\left(m_{kz}, l_q\right) \geq \delta \right) \wedge \left(k = \overline{1..ic} \right) \right\}$,

where: – $f\left(m_{kz}, l_q\right)$ is a function which is defined as $f\left(m_{kz}, l_q\right) = y\left(text_{kz}, WF_q\right)$;

– δ is a threshold of presence of keywords in text messages of current user. It can be defined by decision maker.

$y\left(text_{kz}, WF_q\right)$ is a function of fuzzy search matching, conceptually it can be presented as follows: $y\left(text_{kz}, WF_q\right) = \sum_{i=1}^{g} \sum_{j=1}^{r} d(w_{zi}^k, wf_{qj})$, $d(w_{zi}^k, wf_{qj}) \leq \varphi$,

where: – $d(w_{zi}^k, wf_{qj})$ is a distance measure, which shows similarity between two words w_{zi}^k and wf_{qj}. Initial states are: $d(0, wf_{qj}) = wf_{qj}$ and $d(w_{zi}^k, 0) = w_{zi}^k$,

– φ is a threshold of distance measure, it can be defined by decision maker. The less is the value of distance measure, the higher is similarity between words. That means that current word in text message is a keyword written with mistakes with great probability.

3. Using model of OSNs in automated monitoring system

Automated monitoring system included the following main subsystems:

- data collection which includes crawlers;

- fuzzy text search which includes linguistic knowledge base, keywords database and algorithmic search modules;

- results processing which includes clustering and report generation modules;

- database of text messages and database of search index.

According to model, data collection subsystem gathers identifiers and text messages with additional attributes like type of messages, time and date of posting. This information is stored in database of text messages. Decision maker can specify settings of OSNs crawl strategy.

Subsystem of fuzzy text search takes information from database of text messages and implements the goal of automated monitoring system, trying to detect illegal behavior by using linguistic knowledge base and keywords database. The use of linguistic knowledge base helps to make information retrieval not so sensitive to mistakes. Linguistic knowledge base contains information about inflection paradigms, models of mistakes, typos. Keywords database stores grammatical, semantic information and word forms of keywords lexemes as it was mentioned in the previous section. In case some message contains threshold amount of keywords, this message is indexed and is sent to database of search index. Processes of gathering information by data collection subsystem and searching by subsystem of fuzzy text search are parallel.

According to decision maker's settings subsystem of results processing generates reports and allows viewing information in different slices such as dividing users into risk groups in relation to searching topic.

4. Results and discussions

At the present time automated monitoring system is to be used in detection of drugs use propaganda and illicit drug sales in OSNs [6], though system can be used in different fields, it depends on keywords database. Linguistic database of keywords used in the field of illicit traffic of narcotic drugs and psychotropic substances was developed [7]. Corpus of text messages is gathered from OSN Vkontakte.

Currently algorithms of fuzzy search and models for linguistic knowledge base are developed. Features of algorithms and default values for distance measure should be tested on text corpus and corrected in case of need as they are a kind of empirical data because natural language is not a good formalized object [8, 9]. Model based on hidden Markov model was developed for solving problems of text obfuscation [10]. Users connected with illegal fields of activity may obfuscate their text messages to prevent effective monitoring and hide their actions. For example, letters "o" may be replaced for digit 0, symbols of punctuation may be put among letters, etc. Models for linguistic knowledge base should be compared and chosen the best ones.

5. Conclusion

For providing information and psychological security of users, it is necessary to organize online social networks monitoring. Monitoring process has many difficulties like short messages in OSNs, informal communications and using jargon. Thus, in OSNs modeling emphasis should be on text messages, corresponding model was presented in this paper. Main subsystems of automated monitoring system using aspects of the model were described. Features of future work were given.

References

[1] Davydova, Yu.V To the issue of need for automation of threats search process in virtual social networks and communities / Yu.V. Davydova // Actual problems in modern science in XXI century: proceedings of the 6th international scientific-practical conference. – Makhachkala: "Aprobaciya" Publisher, 2014. – P. 25-26. – (in Russian)

[2] The top 25 social media monitoring tools [Electronic resource]. — Access mode: http://keyhole.co/blog/the-top-25-social-media-monitoring-tools/ (19.01.2017)

[3] Savva, Yu.B. About the problem of the linguistic analysis of the slang in the problem of the automated search of threats of spread of drug addiction on virtual social networks / Yu.B. Savva, V.T. Eryomenko, Yu.V. Davydova // Information systems and Technologies. – 2015. – Vol. 6, no. 92. – P. 68-75. – (in Russian)

[4] Gubanov, D.A. Online social networks: models of information influence, control and confrontation / D.A. Gubanov, D.A. Novikov, A.G. Chhartishvili Moscow: "Fizmatlit" Publisher, 2010. – 228 p. – (in Russian).

[5]Savva, Yu.B. Linguistic database for monitoring system of online social networks in providing information and psychological security / Yu.B. Savva, Yu.B. Davydova // European integration: justice, freedom and security: proceedings of VII scientific and professional conference with international participation: in 3 volumes. – Belgrade: “Criminalistic-Police Academy” Publisher, 2016. – V. 1. – P. 145-154.
[6] Savva, Yu.B. Design of information system identification of persons which participate illicit in field of narcotic drugs and psychotropic substances in the virtual social networks using the database jargon / Yu.B. Savva, V.T. Eryomenko, Yu.V. Davydova // Information systems and Technologies. – 2016. – Vol. 1, no. 93. – P. 68-75. – (in Russian)
[7] Certificate of state registration database no. 2016620197. Jargon in the field of illicit traffic of narcotic drugs and psychotropic substances / Yu.B. Savva, Yu.B. Davydova. – registered 10 February 2016
[8] Ingersoll, G.S. Taming text. How to find, organize and manipulate it / Grant G. Ingersoll, Thomas S. Morton, Andrew L. Farris – NY: Manning Publications Co., 2013. – 320 p.
[9] Manning, C. D Introduction to information retrieval / C.D. Manning, P. Raghavan, H. Schutze – Cambridge: Cambridge University Press, 2008. – 496 p.
[10] Nikol’skaya, A.N. About the problem of opening of obfuscated Russian-language texts of participants of online social networks / A.N. Nikol’skaya, Yu. B. Savva // Information systems and Technologies. – 2016. – Vol. 6, no. 98. – P. 44-55. – (in Russian)