UDC 004.891.3 **DOI:10.18413/2518-1092-2016-1-4-55-63**

Savva Yu.B.

MONITORING AND ANALYSIS OF ILLEGAL ACTIVITY OF PARTICIPANTS OF ONLINE SOCIAL NETWORKS IN THE CONTEXT OF THE SOLUTION OF THE PROBLEM OF SAFETY OF THE PERSONALITY AND THE STATE

Orel State University named after I.S. Turgenev, 95 Komsomol'skaya St., Orel, 302026, Russia
e-mail: su_fio@mail.ru

Abstract
Advances in information and communication technologies have led to a global information society and the emergence of a new digital world in which the development of each individual, communities, states, politics, and economics greatly depend on the use of telecommunications. At the same time global informatization allows to use new information and communication technologies to destabilize social situation in different countries, government authorities as well as to conduct illegal and destructive activities of political, criminal and terrorist organizations.
In conditions of increasing instability of the situation in international relations and global economy there appear new challenges and threats to national security and sustainable development in Russia and other countries, including in the virtual environment. In this regard, the special relevance of the problem of cybersecurity for public authorities, economic entities and ordinary citizens in the processes of their interaction on the Internet, including ONS.
The article puts forward a method of analysis of activity and identification of participants of the online social networks conducting illegal activity. Besides, it considers the structure of the information system of monitoring of their behavior.
Keywords: Cybersecurity; illegal activity; online social networks; telecommunication; information system.

УДК 004.891.3

Савва Ю.Б.

МОНИТОРИНГ И АНАЛИЗ ПРОТИВОПРАВНОЙ ДЕЯТЕЛЬНОСТИ УЧАСТНИКОВ ВИРТУАЛЬНЫХ СОЦИАЛЬНЫХ СЕТЕЙ В КОНТЕКСТЕ РЕШЕНИЯ ЗАДАЧИ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ ЛИЧНОСТИ И ГОСУДАРСТВА

Орловский государственный университет имени И.С Тургенева, ул. Комсомольская, 95
г. Орел, 302026, Россия
e-mail: su_fio@mail.ru

Аннотация
Достижения в области информационных и коммуникационных технологий привели к созданию глобального информационного общества и появлением нового цифрового мира, в котором развитие каждого человека, общества, государства, политики и экономики в значительной степени зависят от использования телекоммуникаций. В то же время глобальная информатизация позволяет использовать новые информационные и коммуникационные технологии, чтобы дестабилизировать социальную ситуацию в разных странах, правительственные органы, а также проводить незаконные и деструктивной деятельности политических, криминальных и террористических организаций.
В условиях возрастающей нестабильности ситуации в международных отношениях и мировой экономики возникают новые вызовы и угрозы национальной безопасности и устойчивого развития в России и других странах, в том числе и в виртуальной среде. В связи с этим, особую актуальность приобрели проблемы обеспечения кибербезопасности для органов государственной власти, хозяйствующих субъектов и простых граждан в

процессах их взаимодействия в сети Интернет, в том числе в виртуальных социальных сетях.

В статье предложен метод анализа деятельности и идентификации участников виртуальных социальных сетей, осуществляющих противоправную деятельность, а также структуру информационной системы мониторинга их поведения.

Ключевые слова: кибербезопасность; противоправная деятельность; виртуальные социальные сети; телекоммуникации; информационная система.

INTRODUCTION

The Internet provides opportunities to enhance the degree of social integration, and at the same time, it contributes to the development of deviant behavior of users due to their partial or complete anonymity. Users can be both active participants in social networking, concerned with promoting and implementing the criminal actions and passive participants, who are exposed to the effects of the relevant information. At the same time the state security threats and identity are the propaganda of terrorism, extremism, drug abuse and drug trafficking through the web-sites of the Internet and online social networks (ONS, web 2.0). These circumstances require monitoring, and in some cases, close monitoring of the user's network in order to analyze their actions and to identify actors' destructive effects on law-abiding participants in social networking.

The VKontakte (InContact) network is one of the most popular ONS in Russia. Being popular, it has led a network of young people whose consciousness is still being formed and to the choice of the VKontakte network to analyze activity of its participants conducting illegal and destructive actions. Quick updates and dissemination of information in social networks makes the process of monitoring and analyzing reports difficult. Furthermore social networks represent large amounts of data, which are practically not indexed by traditional search engines. In addition, the message texts are not always structured and grammatically correct. To complicate the understanding of the topics discussed by the uninitiated, to complicate the search and discovery of persons involved in criminal activity, the communication often uses a special slang.

To solve the problem of activity analysis and identification of participants in the network VKontakte there has been developed an information monitoring system of the network. The structure of information for monitoring and analyzing network system includes 6 software packages and four data bases.

1. THE SOFTWARE PACKAGE TO IDENTIFY THE ACTIVITY OF PARTICIPANTS IN ONS

This software package is designed for the formation of the history of the activity of selected participants in ONS, including date and time of the activity status and the device, including its type, with which there were recorded the visits to personal pages of participants. Gathering participants' activity is carried out using a «cron» type task scheduler on the server every 5 minutes that runs a script that collects the required data, which are entered into a database. A feature of this software package is to use the structure as an architectural pattern MVC (Model, View, Controller), and the presence of an expandable list of device types that are used by users to visit the ONS. There was developed a software package to identify the activity of the participants of ONS for desktops (Windows XP and above) as well as for mobile devices and tablets (running Android 4 and above).

The result of the operation of software package for identifying activity of the controlled participant of ONS is presented in figure 1.

2. SOFTWARE PACKAGE FOR AUTOMATED CONSTRUCTING OF SOCIAL GRAPH OF CONTACTS OF PARTICIPANTS OF ONS

This software is intended for information search about contacts of a particular participant's contacts in ONS, having a personal identifier, identifying connections between a pluralities of found contacts and constructing a social graph. In terms of data mining, ONS is a heterogeneous multirelational array of data presented in a graph. Therefore, to study the ONS structure in this software we used the cluster analysis algorithms to split the network nodes representing objects into classes based on their relationships as well as their attributes.

Information search is conducted for keywords and phrases in the pages of the participants of ONS, as well as on the "Wall" – the data is laid out by these parties on the overall review and comments.

Fig. 1. Report on the activity of the controlled participant in ONS VKontakte

The object of the database "Participants", which describes the participants of social network VKontakte, has the following structure:

1) Standard fields

−id (identifier),

−first_name (name),

−last_name (surname),

−deactivated (refundable in case if the user page deleted),

−hidden (returns the unit if the user page is hidden from outsiders);

2) Additional fields including:

−nickname,

−activities,

−occupation,

−status,

−last_seen,

−followers_count,

−common_count,

−wall_comments,

−lists (list IDs of friends) and others.

A total of 46 additional fields provided.

A feature of this software package is incorporated in its base graph drawing an algorithm based on particle physics with gravity field around each node, and the links mechanism is implemented on the basis of the springs.

Figure 2 shows a graph of contacts of the controlled participant's ONS.

3. CRAWLER

Crawler collects data by iterating through the pages of users and communities in ONS. Content reviewed by the crawler pages is passed to the indexer.

Depending on the problem to be solved by the user of information system fuzzy text search in the text of the message of participants of ONS can be carried out in the following variants (for example, an evolved search for participants, promoting the use of narcotic drugs and psychotropic substances):

−Using the entire database jargon;

−Only for certain selected groups of illegal drugs and psychotropic substances;

−For the jargon related to the group of the most characteristic of the speech and signaling addicts on unique accessories authors of these messages to the field of drug addiction.

Indexer is a collection of several modules where the input data for one module are output to another. The content of the page in question is subject to an initial analysis of the natural text presented in the form of a chain of ASCII characters, generates the information needed for further processing morphological and syntactic processor, the result of which is a set of lexical units. At this stage, if necessary, a word with a certain degree of probability exposed deobfuscation subject to the rules remove the obfuscation. Next fuzzy text search is carried out among the set of lexical units with the use of a linguistic database (LBD) jargon. The database accumulates slang and word forms permissible in a particular area. Accounting morphology will increase the accuracy of the information retrieval. In addition, there is interpretation and storing for each rank jargon. Rank is the degree of specialization jargon -

vernacular vocabulary, slang and vernacular specialized vocabulary. Vernacular vocabulary becomes jargon only in a context that is particularly special difficulty in identifying keywords in the text. Common words form a collocation - phrases that are signs of syntactically and semantically coherent units. This collocation receive new semantic value is not peculiar to their lexical components separately.

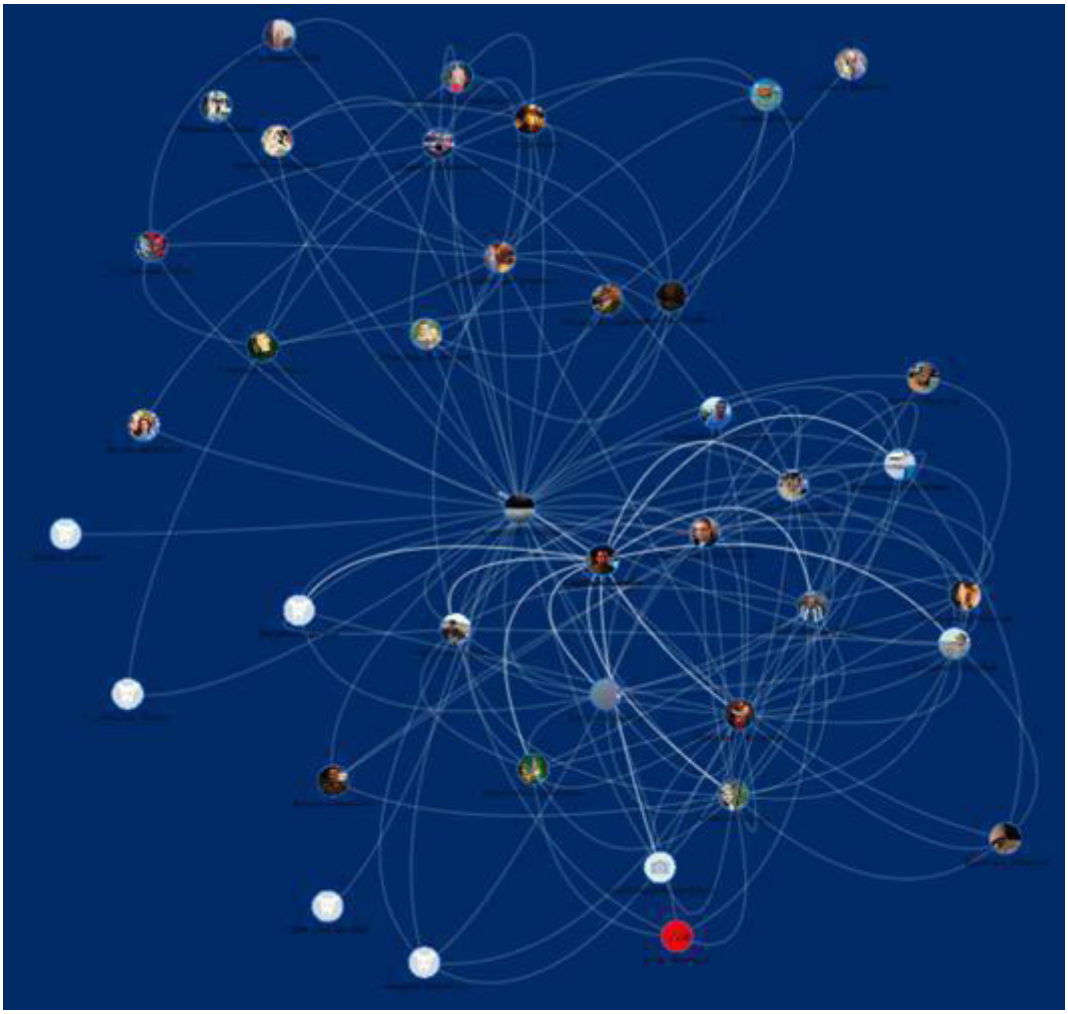

Fig. 2. Graf of contacts of the controlled participant of an ONS

If the page content contains keywords found with the degree of probability, given at the fuzzy text searching, then the page is indexed. Thus, the index database stores information only about the suspicious pages. Unlike traditional search engine, the ranking in the monitoring system of the results takes place before the formation of the index. Such an approach is determined by the purpose of the monitoring system – detection of individuals involved in the promotion of deviant behavior and provocations. Pages containing highly specialized words, that are located on the top search, are already in the process of forming the index.

4. PACKAGE FOR REMOVING OBFUSCATION OF THE TEXTS

The obfuscation of the texts in ONS may be applied to conceal the machinery from the search correspondence concerning illegal activities such as trafficking of drugs or weapons.

In general, the following methods are used for obfuscating the text message (either separately or together): intentional misspelling; improper segmentation of words; substitution alphabet to characters similar in appearance or character combinations; pasting insignificant characters; for the Russian language – the use of transliteration. It is necessary to take into account that the participants of ONS often use slang and neologisms.

Most of the existing research is focused on content filtering and spam e-mails. The most commonly used (alone or in combination) include: regular expressions, N-programmatic probabilistic models, phonetic similarity search obfuscated strings with key words, the distance Damerau-Lowenstein. However, these methods do not solve the problem of opening the obfuscation of texts for all kinds of deobfuscation. The method, which allows to solve the problem of deobfuscation in all these ways earlier text obfuscation, is the use of a hidden Markov model (HMM). This probabilistic model code obfuscated text is observation, and the task is to find a sequence of hidden states corresponding to the characters of the alphabet, which most likely correspond to derived observations. This package uses the Viterbi algorithm.

This model uses the lexical tree constructed on the basis of a dictionary and determining the probability of transition between states HMM (tree nodes), based on pre-calculated word frequency. For the processing of words that are not in the dictionary, there is a second part of the model. As the transition probabilities between the states of the second part of the model used by the statistical data of letters to each other for a specific language. The probability of observing symbols when in a particular hidden state is empirically determined from training patterns.

A disadvantage of this model is the large number of states, which constitute the lexical tree that leads to considerable working time of the Viterbi algorithm. This issue is critical in the case of the Russian language: in it, unlike English, which was developed for the specified model, significantly there are more word forms for each word. Each word must be stored in the tree for the accuracy of deobfuscation, which increases the number of states (instead of word forms in the lexical tree may store only word stems to form a word form under the rules of the language, but for jargon and neologisms detected during the operation, the automatic detection of word forms is not trivial, so that the construction of tree using word forms) was chosen.

To solve this problem, all of the state's corresponding to one symbol alphabet determined in the super-state. For each super-state is entered redundant, allowing to determine the cases forcing the insertion of characters and repeating characters. The transition probabilities are determined by checking whether it is possible to find obtained in this step of the Viterbi algorithm a sequence of hidden states in the lexical tree, considering only that part of the sequence that corresponds to the intended last word.

The package uses two approaches: the use of super-states of the lexical tree and states for the words out of the dictionary (also duplicated to account for the possibility of inserting characters). This algorithm has been developed that allows the lexical tree be left to stand too long in view of the addition of new words. This allows you to more accurately conduct deobfuscation, identifying and adding the most common words and adding specialized dictionaries on particular subjects (eg, substance abuse). Basic Dictionary was compiled based on the most frequently used word forms of the Russian language, which cover about 70% of the text, "The Russian National Corpus."

For correct operation of the HMM required parameters defining the probability that in a latent state does not occur surveillance symbol (delete character in the word), a transition to the redundant hidden state (symbol insertion), the probability distribution of the symbols of observations for each super-state, and the ratio which determines the ratio of the transition probability in one of the two parts of the HMM. To calculate these parameters using the machine learning is used sampling obfuscated messages. To calculate the frequency of words that do not belong to the initial dictionary used accumulated in the process of bringing the frequency statistics from the word to the unit «imp» (number of cases per million words).

5. STATISTICAL TEXT ANALYSIS PACKAGE

The purpose of this package – the collection, evaluation and optimization of the statistical data of the Russian language for the implementation and effective functioning of the statistical language models within the obfuscation algorithm opening message texts.

Since the implementation of the algorithm opening obfuscation as a data representation model uses a HMM, it is necessary to calculate its parameters, in particular, such as the probability distribution of observed states. This problem is solved as follows. It is necessary to create a table of possible obfuscation (the observed states of the model) of the respective characters of the Russian alphabet (hidden states of the model). Next, you need to optimize the initial assumptions about the parameters of the model based on the marked training sample. As an optimization method is used EM-algorithm for the maximum likelihood estimation.

6. PACKAGE FOR MESSAGE CLASSIFICATION AND CLUSTERING OF SOCIAL GRAPH

Classification of participants in the social network on the data from their posts held by

constructing mathematical model messaging of participants of ONS. This allows the message text classification (formally, to determine the category of simulated), to make recommendations on how to add new terms to the dictionary with an indication of their possible categories, as well as the classification of participants of ONS based on the categories of messages. On the basis of the text messages housing members of the network, methods of topic model generated a thematic model that determines the probability distribution of the documents and the probability distribution of words in the topics. As a method of thematic modeling algorithm used latent Dirichlet allocation (LDA).

Next to the distributions of words each topic classification method used to determine the topics in a particular category. For this training set is used, which is actually a dictionary where each word is marked categories. After that, the decision on affiliation of each user to a certain class of pre-characterizing suspicious message content network participant. For the above two actions appropriate to use a naive Bayesian classifier as the implementation of the method. It is also one of the results of the thematic modeling is the ability to identify new words not found in the dictionary, and a recommendation on awarding them to the category.

The clusters highlighted on the graph, which is shown in figure 2, are presented in figure 3.

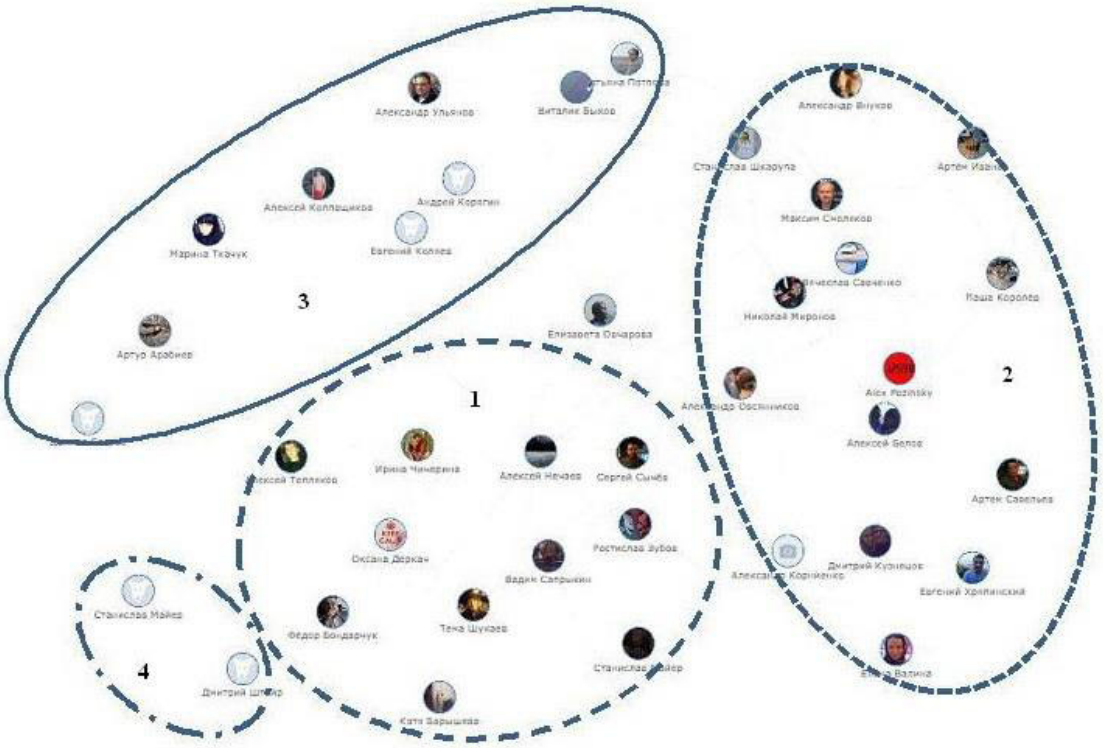

Fig. 3. Clustering of the graph of contacts

The clusters presented in figure 3 is selected according to the frequency of contacts: 1 – weekly contact; 2 – contacts with periodicity not less often than once a month; 3 – contacts, carried out not more frequently than once a month; 4 – contacts, carried out with a frequency of less than once in three months.

7. DATABASE

The database of information system may contain various terms for text search. Consider a database structure in the example of the jargon which is used in the sphere of illicit trafficking in narcotic drugs and psychotropic substances.

To identify the jargon of the database in the text of the message we need to solve the problem which is called the identification of entities or recognition of named entities. As a result of solving this task were allocated to the following entities:

–Substance – narcotic drug or psychotropic substance;

–SubstanceGroups – group of narcotic drugs and psychotropic substances;

–Semantics – entity denoting semantics of the jargon;

– WordFroms – word form;

– PseudoBasis – pseudosasa of jargon.

In figure 4 the fragment of the chart of classes of information system of search of jargons in the sphere of illicit trafficking in drugs and psychotropic substances is presented [1].

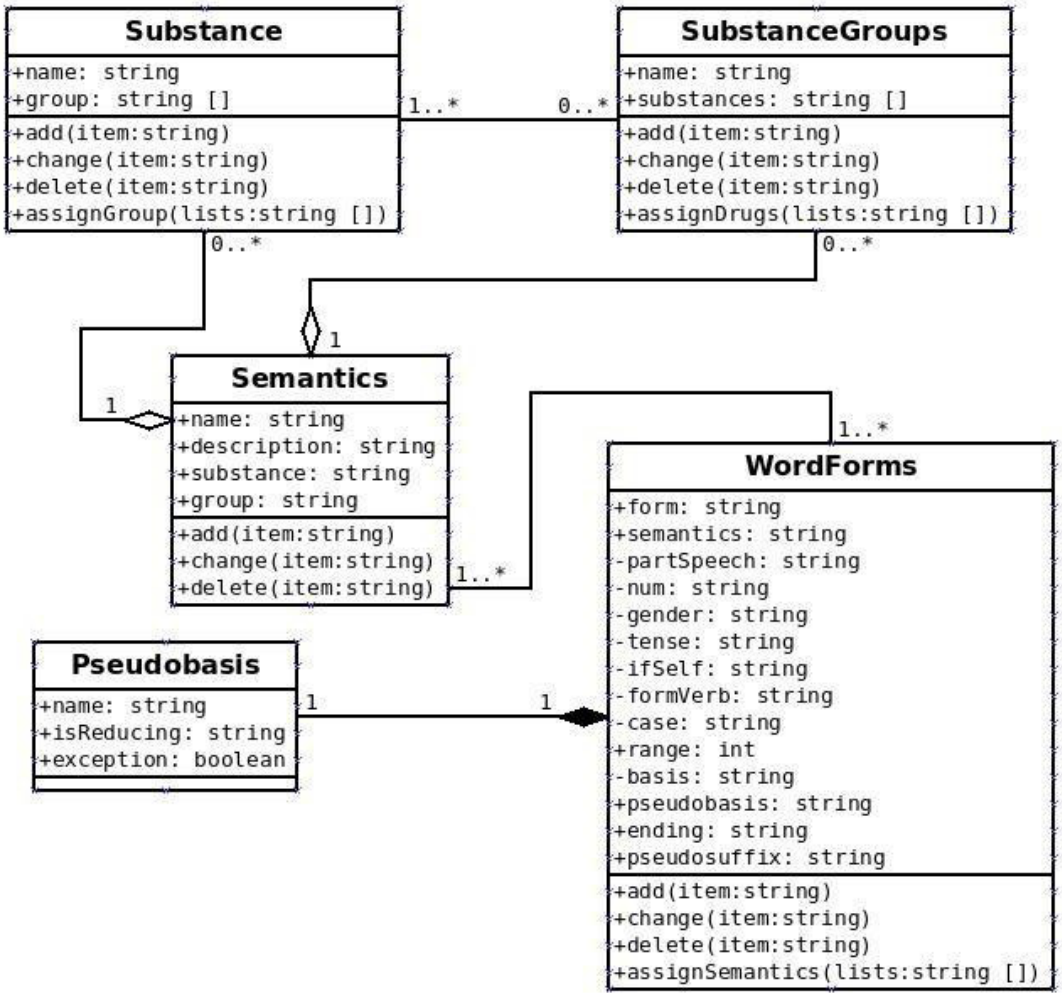

Fig. 4. Fragment of the chart of classes of information system of search of jargons in the sphere of illicit trafficking of drugs and psychotropic substances

Drugs are mostly distributed in groups of substance (assignGroup method), but allowed to have a database of substances not associated with any of the existing groups. Methods “add”, “change”, “delete” allow us to add a new drug, edit and delete an existing drug substance, respectively.

The user of information system has a possibility of creation of new groups of narcotic substances according to any desirable criterion, for example, the drugs limited in a turn or new synthetic drugs. Then substances from the database are brought in new group.

Semantics of jargons can represent the name of drug or the whole group of substances, and also designate the narcotic state, tools, persons distributing and using drugs, etc. There is a probability of a polysemy when one jargon can designate several values and vice versa, one type of semantics can extend to a set of jargons. The “description” attribute of essence “Semantics” allows the user to add the expanded text description for this or that value of a jargon by the principle of the explanatory dictionary.

For word forms pseudo-bases which are used by information search are allocated. The ending and pseudosuffix attributes of essence WordForms represent the termination and a pseudo-suffix of a word form respectively. The range attribute defines degree of a specialization of this or that word form, basis represents a word basis if at word change the reduction fact takes place, other attributes are grammatical features. The assignSemantics method allows appropriating to a word form a set of values.

CONCLUSION

Linguistic analysis of text messages the participants of ONS using a database developed by the jargon in the field of illicit trafficking in narcotic drugs and psychotropic substances allows using software to conduct the construction and analysis of the structure of these networks, as well as the properties of individual communications and thus solve the problem of identification of actors included in the promoting individual and community distributing narcotic drugs and psychotropic substances on the Internet.

The information system considered in this paper is used in regional departments of Russian police.

Currently underway database work to fill these terms from other areas: terrorism, extremism and others. This will extend the scope of the information system for the monitoring of ONS to solve the problems of safety of the personality and the state.

References

1. Savva Yu.B., Davydova Ju. V. Linguistic Database for Monitoring a System of Online Social Networks in Providing Information and Psychological Security // Collection of works of the VIIth Conference “European integration: justice, freedom and security (2016, Tara)”, V. 1.: Academy of criminalistics and police studies of Republic of Serbia, Belgrade, 2016. Pp. 145-154.

Савва Юрий Болеславович, кандидат технических наук, доцент кафедры информационной безопасности

Yuri B. Savva, PhD in Technical Sciences, Associate Professor, Department of Information Security