Analysis of the personal information from social networks to solve the problems of criminology

E A Gambarova[1], V A Bakaev[1], N V Olinder[1], A V Blagov[1] and M E Naumov[1]

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

Abstract. The article discusses the need to use social networks in the cognitive activities of participants in the criminal process, and suggests that it is possible to use information obtained from social networks in the investigation of crimes. Two approaches are compared: expert and automated. The authors offer tools for data collection and analyzing personal data from social networks.

1. Introduction

An important condition for improving the effectiveness of combating modern crime is the continuous improvement of theoretical and practical knowledge of the investigator, investigator and operative worker on the use of modern technologies in the investigation of crimes, as well as the search for new ways to collect information.

For example, recently much attention has been paid to technical systems for face recognition [1,2,3].

The information is central to the cognitive activity of the investigator, so the search for ways to obtain the information more quickly and fully is an important area in criminology. Great opportunities for working with information are provided by the Internet, in particular, social media (social networks.

At present, it is necessary to note the change in the approach of citizens to the methods and forms of communication, information circulation, etc., partly due to the development of the global Internet, the development of virtual relations. Communication via social networks and messengers is gaining more and more momentum [4]. The growing popularity of social media (social networking, instant messengers, etc.) [5], the spread of "virtual" databases, online banking, cloud storage and other tools used for more comfortable and fast communication and receipt (supply) of services leads not only to the need for normative regulation of these relations, but also determines the creation of new approaches to virtual space in criminology.

The need to use information from social networks in the investigation says the current investigator on particularly important cases of the Republic of Belarus, K. Yu.n. Yu. F. Kamenetsky [6]. In his dissertation "the methodology of the initial phase of the investigation of embezzlement by abuse of official authority in the public sector", he writes that "Anticipation of action by the embezzler and his corrupt ties directly linked to the analytical work to establish social and family ties to the looting, the adoption of measures of preventive character. One of the effective ways of investigating the investigator's activities, along with the classical measures, is the monitoring of social networks available on the global computer network.

This problem is typical not only for forensic and procedural science in the Russian Federation, but also for other countries, as "virtualization" is an integral part of the globalization process taking place

around the world. According to Professor Volchetskaya T. S., the problems associated with virtual space are particularly promising for scientific development. So some steps to forensic and procedural knowledge of the process of virtualization for several years, Russian scientists Efimov V. Yu, Vekhov V. B., Volchetskaya.S., Ishin A. M., Meshcheryakov V. A., Olander N. In. Smushkin A. B. etc. So, Meshcheryakov V. A. one of the first in his work "bases of a technique of investigation of crimes in computer information sphere" referred to "virtual tracks». Subsequently, the section "forensic study of traces" was supplemented with scientific developments related to the study of nature, nature, species, processes of formation, identification and consolidation of virtual traces in the conduct of individual investigations.

When planning certain investigative actions, the investigator is tasked with choosing the most effective ways to achieve this goal. As a rule, one of the tasks of planning investigative actions is to collect information that will help the investigator to choose the tactics of this or that investigative action in the future. Due to the fact that the investigator does not always have a large amount of time to search for information about the event of interest or personality, it is necessary to choose ways that will help to reduce the time of obtaining information. It seems promising to use information from the Internet, including the monitoring of social networks, when planning investigative actions, in particular, interrogation.

2. Methods of social network data collection and processing

In criminology, the organization of the investigation is necessary to quickly obtain reliable information about a person or group of people. Efficiency, reliability and timeliness are key factors. Therefore, it seems reasonable to develop the most effective technology of obtaining and processing information, thus, present the necessary methodological and methodical issues related to the application of the tools of the Internet (social networking, etc.), elaboration of algorithm of search and verification of information. The process of constructing an algorithm for solving a professional problem, the result of which is the allocation of stages of the data processing process, the formal definition of their content and the order of their execution, the development of a template of actions and/or mental operations, allows to optimize the search-cognitive, organizational and technological component of the active interaction of the person conducting the investigation with the objects of the world associated with a criminal event [7].

The task of collecting the necessary information from social networks can be divided into data collection, filtering, processing and subsequent analysis.

The authors of the study set the task of collecting data of social networks expertly, as well as using the developed software. In the first case, a group of experts was determined who systematically searched for the necessary personal information without the use of additional automated services. In the second case, the following approach was used.

Based on the task, the developed software package implements the following functionality:

- analysis of all profiles of target social networks (Vkontakte, Twitter, Instagram, LinkedIn) in order to save open information in the database;
- matching the profiles belonging to one person in the group;
- making assumptions about the user's income level.

For the implementation of a software product, we used the following technology stack: Scala, Python, PostgreSQL, ApacheStorm, CatBoost. This choice is due to the requirement for horizontal scaling of the system.

CatBoost is used to build a mathematical model that determines a person's income level by such parameters as: gender, age, education, field of activity, position, city, family status.

To search for a person's accounts in other social networks, we use a two-layer perseptron, comparing profiles by name, nickname, email, etc.

Primary data are collected as follows [8].

At the first stage the system analyzes all users of social networks VKontakte, Twitter and Instagram and groups them in the following rules:

- in each group there are no more than one profile from each social network;

- all profiles in one group belong to one person.

This problem is solved by means of a program framework of Apache Spark (in particular, superstructures of Spark Streaming intended for stream data processing see fig. 1) and the broker of messages RabbitMQ realizing delivery of basic data in Spark Streaming.

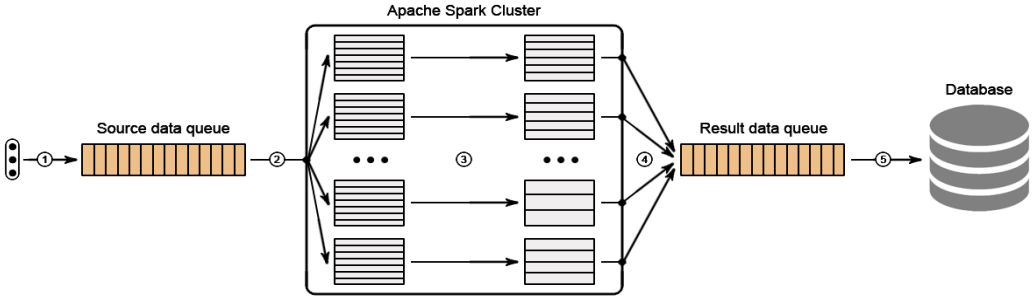

Figure 1. Architecture of the aggregator of users of social networks (pipeline).

Description of steps:

1. Adding of data from different sources in queue for later processing. Data represent a set of couples (network_id, user_id) containing information on profiles which are required to be analyzed.
2. RDD (Resilient Distributed Dataset) formation by a packing of the basic data which are in queue for increase in productivity.
3. RDD (mapping) conversion. For each couple (network_id, user_id) the algorithm finds and groups profiles on other social networks, and also additional information on the person to whom belongs the initial account. The algorithm is restarted for each found profile until all available information on the user is found. As sources can be: the public information specified on the page of the user (the status, contact information, entries in the film, etc.).
4. Export of data retrieveds from RDD in queue for the subsequent saving.
5. Saving results in NoSQL to the MongoDB database in the form of documents with structure, the reflected in table 1.

The speed of data processing makes about 120-130 profiles a second. For work the Microsoft Azure A2 v2 virtual computer was used (2 kernels, 4 GB of RAM, 20 GB of SSD). Casual users of social network VKontakte (1.000.000 profiles) were analyzed.

Thus, if to assume that speeds of processing of profiles of VKontakte, Instagram and Twitter are equal, we will receive an approximate assessment of time which will be required for the analysis of all users of target social networks:

$$T(n) = \frac{4*10^8+6*10^8+13*10^8}{120n} \approx 5324hours \approx 221day, \quad (1)$$

where n - the number of servers in a cluster with a similar configuration.

In case of horizontal scaling of a cluster the linear dependence between the number of servers and processing rate of profiles is watched.

Example of the reference to the table: results of an experiment are reflected in table 1.

Table 1. Data storage structure of profiles.

Name of the field	Type	Description
_id	ObjectId	the document identifier in a collection
vk_id	Int32	the profile identifier in VKontake
facebook_id	Int64	the profile identifier in Facebook network
instagram_id	Int64	the profile identifier in Instagram network
twitter_id	Int64	the profile identifier in Twitter network
other	Object	The additional information (phone number, e-mail address, skype, etc.)

The further task comes down to expansion of the received base by association of profiles by the rules described earlier on which pages bek-links aren't specified other social networks.

Data grouping is based on the analysis of common features [8]. The following procedure applies:

Construction of a full multi-column graph, which stores information about the profiles of social networks and the potential that characterizes the probability of their belonging to one person.

At the vertices of a graph contains information about the profile that is used when the comparison is made.

To compare two profiles, a multilayer neural network is used. The input network layer is fed with a vector of dimension 12, which contains the following data:

Name ↔ Name'
max(Name → Username', Name' → Username)
max(Name → E-mail', Name' → E-mail)
max(Name → Skype', Name' → Skype)
Username ↔ Username'
max(Username → E-mail', Username' → E-mail)
Username ↔ Skype'
max(Skype → Username', Skype' → Username)
max(Skype → E-mail', Skype' → E-mail)
E-mail ↔ E-mail'
Phone ↔ Phone'
Website ↔ Website'

Next, it defines the fullness of occurrences of a in b:

$$a \to b = 1 - \frac{d+r+s}{len(a)} \in [0,1], \quad (2)$$

where d – is the number of delete operations to convert a to b; r – number of replacement operations to convert a to b; s - number of transposition operations to convert a to b; len(x) – function to calculate the length of the argument.

Comparison a and b:

$$\forall i \in [1, len(a)], j \in [1, len(b)] d[i,j] = 1 - \frac{dist(a[i],b[j])}{len(b[j])} \in [0,1], \quad (3)$$

$$a \leftrightarrow b = \frac{\sum_1^{len(a)} d[i, fit(i)]}{min(len(a),len(b))} \in [0,1],$$

where dist(a, b) - is a function that calculates the Damerau-Levenshtein distance [9] for lines a and b; fit(i) - a function that returns the index of the word of the string b, put in accordance with the word a[i].

The comparison operation does not consider the word order. All the words in the source strings are compared in pairs, and then, using the algorithm of Kuhn-Munkres [10], each word of a string a is defined in accordance with the word line b so that the sum of the similarity for all pairs of words was the maximum. Also, punctuation marks and other symbols (except letters and numbers) are not taken into account.

Training and control samples are collected on the basis of primary data. The size of the training sample ~106 pairs.

Next in the generated graph for each pair of shares the following sequence of actions is performed:

- edges are sorted in descending order of weights;

-edges whose weight is less than the threshold value are removed, or one of the incident vertices is already connected to some vertex of the opposite fraction.

The result of these transformations is a graph in which each component of connectivity is a group of accounts from different social networks that belong to one person.

Due to the fact that a person can belong to several communities at the same time, and if the same group was formed in several communities, it can be assumed that the accounts of this group really belong to one user.

Regular expressions and Scala's built-in mechanism of working with CS-grammars are used to parse contact information.

3. Results and discussions

Using the method of expert data collection and processing of social networks, the following results were obtained. The following experiment was conducted. The expert group (115 people), offered to look for information about certain people (3 people) on the given parameters: the place of residence of the person, the place of study of the person, joining the founders, the presence of property, the presence of debts and fines, participation in trials, travel and business trips, leisure, family, close friends.

An important condition was that the participants of the experiment did not use special technical means and looked for information only in open sources.

As a result of the experiment, it was found that it is easy to find information about the city of residence (67% of participants found), although the specific address was found only 5.7% of participants. 79% of the participants were able to find the place of work and study. Joining the founders, shareholders, the status of an individual entrepreneur, etc.found 22% of the participants of the experiment, the availability of information about the property could find only 10% of the participants. Less than 2% of the participants found debts, fines and loans, less than 1% found participation in trials (as a party), more than half of the participants of the experiment could find travel and business trips, 50.6% could find information about parents, 29.5% of the participants of the experiment , about the spouses of more than half of the participants, 52.5%; about brothers and sisters, 19%, about friends, on average 20.9 % (see table 2).

Table 2. TheComparison of two approaches (expert and software).

	The search option	The percentage of participants in the experiment who discovered the information
1	The city of residence	67
2	The residential Address	5,7
3	The place of work and study	79
4	The joining the founders, shareholders, having the status of an individual entrepreneur	22
5	debts, fines, loans	2
6	The fact of participation in court proceedings (as a party)	1
7	Travelling and business trip	50,6
8	The information about parents	29,5
9	The information about spouses	52,5
10	The information about brothers and sisters	19
11	The information about friends	20, 9

The search and processing of information (including the determination of its reliability also took an average of two and a half hours).

Thus, during the experiment, certain parameters ("beacons") were identified, which are freely available in social networks and can be accessed by any user: address, place of residence, place of study, places of rest, business trips, etc.

On the one hand, the experiment showed how much information about the person is stored in social networks, which is a negative factor, as the level of personal data protection is reduced (although these data are placed by the subjects themselves – freely, at will). On the other hand, such "openness" of information can help in the work of investigative bodies in the investigation of crimes. For example, when collecting information about possible participants in organized crime groups, in preparation for individual investigations (interrogation, confrontation) or in General, when planning the investigation of certain types of crimes.

In Russia, the use of social networks in the investigation and prevention of crimes is not widespread. In order to find out the reason of such unpopularity of the use of social networks in the investigation of crimes in the framework of the study, a survey of investigators of the Samara region was conducted. The reasons for this are: the complexity of the search for information (76%), the duration of time (34%), the difficulties of procedural registration of search and use of such information (91 %).

After analyzing the answers of investigators, we concluded that the reduction of time to search for information, as well as the possibility of using some algorithms or software, could create conditions for a wider use of social networks in the investigation of crimes. In this connection, it is advisable to consider the following method of collecting information in the framework of the study.

The second method using the developed software was obtained as follows. The created system of creating portraits of users of social networks is able to collect the following data:

- user profile identifiers in other social networks (including if they are not explicitly specified on their page);
- other contact information (phone numbers, email addresses, Skype logins);
- User's name and nicknames;
- the date of birth;
- city of residence;
- relatives (parents, children, brothers, sisters);
- education (University, school);
- place of work and position;
- уровень дохода (using HeadHunter and Yandex.Work statistic services).

As an experiment, the participants of the community "Big village" were analyzed (https://vk.com/bigvill). For 197 seconds processed data 48.525 profiles Vkontakte (Instagram – 8734, Twitter – 4367, LinkedIn – 1455).

As a result, you can see that with the help of the software you can collect and process much more information more quickly, while, of course, a more detailed analysis can be carried out expertly. The developed software product can be used in criminology for operational preliminary analysis of personal data, including for checking their reliability (for various parameters, for example, according to the specified dates).

Table 3 presents a comparative analysis of the two approaches: with the help of experts and with the help of software.

Table 3. The comparison of two approaches (expert and software).

	Expert	Using the software
The labor expenditures	115experts, 2 hours	The software for 1 PC, 197 seconds
The number of processed data	the social profiles. networks, websites and other open sources	The profiles:VK – 48 525, Instagram – 8 734, Twitter – 4 367, LinkedIn – 1 455.
The completeness of the information	The high, including information about fines and loans	The average: - personal data, - contacts, - connections, - place of work and approximate income level

In General, we can say that social networks can be considered to solve the problems of criminology and serve as the object of research. On the one hand, openness and, as a consequence, availability of data is a negative factor, as the level of personal data protection is reduced (although these data are placed by the subjects themselves). On the other hand, such "openness" can help in the work of investigative bodies in the investigation of crimes. For example, when collecting information about possible participants of criminal groups, when preparing for certain investigative actions (for example, interrogation, confrontation) or in General, when planning the investigation of certain types of crimes.

4. Conclusions

The result of the work is the study and comparison of methods of collection and processing of personal data of users of social networks to solve the problems of criminology. It was found that using social networks you can find a lot of information that users about themselves, their relatives, their work and studies leave on their own, some information is left about users by other users, for example by posting joint photos, videos. These circumstances together make it possible to make a fairly detailed dossier on active users of social networks, which may be important for the investigator, and the information obtained can provide significant assistance in the investigation of crimes.

It is obvious that the comparison of performance and quality of information received by man and machine, gives a predictable result. However, the use of such software by law enforcement agencies can significantly reduce the time to search for basic information and discard from the sample of people who do not meet the specified criteria. And then additional information can be collected by expert.

5. References

[1] Nemirovskiy V B, Stoyanov A K and Goremykina D S 2016 Face recognition based on the proximity measure clustering *Computer Optics* **40(5)** 740-746 DOI: 10.18287/2412-6179-2016-40-5-740-745

[2] Rybintsev A V, Konushin V S and Konushin A S 2015 Consecutive gender and age classification from facial images based on ranked local binary patterns *Computer Optics* **39(5)** 762-770 DOI: 10.18287/0134-2452-2015-39-5-762-769

[3] Nikitin M Yu, Konushin V S and Konushin A S 2017 Neural network model for video-based face recognition with frames quality assessment *Computer Optics* **41(5)** 732-743 DOI: 10.18287/2412-6179-2017-41-5-732-742

[4] Dupuis M, Samreen K and Joyce H 2017 “I Got the Job!”: An exploratory study examining the psychological factors related to status updates on facebook» *Computers in Human Behavior* **73** 132-140

[5] Olinder N V and Gambarova E A 2017 About the results of the experiment "the search and perception of information about a person on the Internet" and its use in the investigation of crimes *Expert-kriminalist* **4** 29-31

[6] Kamenetskiy Yu F 2016 *Methodology of the initial stage of investigation of theft through abuse of official authority in the budgetary sphere: dis. cand. jurid. science* (Minsk) p 196

[7] Stepanenko D A 2016 Algorithmization of the search and cognitive activity of the person conducting the investigation: the grounds, possibilities, problems *Russian Investigator* **10** 3-7

[8] Bakaev V A and Blagov A V 2017 The analysis of profiles on social networks *CEUR Workshop Proceedings* **1903** 88-91

[9] Smetanin N 2011 *Fuzzy search in the text and the dictionary* (Access mode: https://habrahabr.ru/post/114997) (in Russian)

[10] *Hungarian algorithm for solving the assignment problem* (Access mode: http://e-maxx.ru/algo/assignment_hungary) (in Russian)

Acknowledgments

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