An integrated approach to mapping user profiles on social networks

Vladimir Belov
Information Systems department, Faculty of Information Systems and Technologies
Ulyanovsk State Technical University
Ulyanovsk, Russia
belo.199666@mail.ru

Dmitriy Drozdov
Information Systems department, Faculty of Information Systems and Technologies
Ulyanovsk State Technical University
Ulyanovsk, Russia
dimoxx123@gmail.com

Roman Shakurov
Information Systems department, Faculty of Information Systems and Technologies
Ulyanovsk State Technical University
Ulyanovsk, Russia
relife@inbox.ru

Vadim Moshkin
Information Systems department, Faculty of Information Systems and Technologies
Ulyanovsk State Technical University
Ulyanovsk, Russia
v.moshkin@ulstu.ru

Ilya Andreev
Information Systems department, Faculty of Information Systems and Technologies
Ulyanovsk State Technical University
Ulyanovsk, Russia,
ia.andreev@ulstu.ru

***Abstract*—In this paper, we consider an integrated approach for the sole identification of a person in several different social networks by analyzing the questionnaire data, poorly structured information and images comparison from the profiles of social networks. Also the paper contains the description of the software service that implements the proposed approach.**

Keywords—social network, account, search, mapping

I. INTRODUCTION

The active growth of the audience of social networks has led to the emergence of these resources as a new source of data and knowledge. In Russia, several social networks are currently the most popular, each of which has its own focus and specificity of the content posted. Such resources include VKontakte, Odnoklassniki, Instagram, Facebook, Youtube, Twitter[1]. Many users have several accounts on different social networks and publish different or similar content to them. And to find a person in any of the networks becomes problematic.

Working with social networks can be beneficial in implementing the functions of the company's personnel management system, as you can often find out much more information about the professional and personal qualities of the applicant from social networks than from the CV. Currently, the collection or meaningful analysis of information from social networks is carried out manually by specialists of personnel services, which is time consuming and limits the amount of information processed.

Thus, there is a need to develop a software system that allows you to identify a person's profile in several social networks. Such developments would allow aggregating more data about users to assess the severity of their personal characteristics. This work is aimed at solving the problem of searching for an integrated approach for mapping (comparing) user profiles in various social networks based on the analysis of structured data, text information, as well as graphic materials for the purpose of further analysis of the user's social portrait.

II. THE MAIN APPROACHES TO SOLVING THE PROBLEM OF USER IDENTIFICATION

A. Methods and algorithms for mapping user accounts on social networks

Currently, the task of identifying users using data profiles of social networks is solved in various ways [1, 2, 3].

In [4, 5, 6, 7, 8], methods for analyzing data profiles of the social networks MySpace, StudiVZ are described. But these networks are not popular in Russia. The proposed approaches consist in constructing feature vectors of user characteristics based on the information provided on personal pages. To the obtained vectors, methods of exact, partial and fuzzy comparison are applied. In these works, the authors proposed features that are most significant when comparing accounts. The developed algorithms the accuracy of about 80% on a test sample of user accounts.

In [6, 7, 9], methods for mapping user profiles of social networks by analyzing published unstructured (text) information are presented. In [6], the authors conclude that the creator of a text note can be identified by a unique writing style. In [7], a method is shown that takes into account not only text information published by a user in a note, but also meta-information associated with it: geolocation, publication time, hashtags, etc.

B. Software services for searching users in social networks

Currently, there are several services for searching for profiles of people in social networks in RuNet. Most services work on the principle of conventional search engines - download all available open profile data and save it to a local database.

FindFace [10]. One of these services is the FindFace system, as well as many other systems based on it, which allow you to find a person's profile on a social network from their photo. To start the search, you need to select a photo where the human face is clearly visible, and upload the picture. The algorithm will find pages with similar photos and lay out links to them with examples of images. Each link will have a rating from 0 to 1. If the indicator is more than 0.67, then this means that the system recorded the most complete match. The developed neural network scanned the faces of 500 million users of the VKontakte social network.

Yandex.People [11]. The Yandex.People system uses text data obtained from social network profiles to search. So the following data is uploaded from the profiles of a person:

- Name of the user (or at least one of the parameters allowing to identify the person).
- Age of user.
- Place of residence or user address.
- Place of study or completed education.
- User's place of work.

Despite the availability of these services that solve specific tasks of searching for users of social networks, there are currently no comprehensive approaches and universal services that allow users to compare user profiles in various social networks by analyzing not only the data of profiles, but also poorly structured information from the pages of the respective accounts.

III. Approach to Mapping User Profiles on Social Networks

The developed algorithm takes as a basis a person's personal page from a social network. Different information is downloaded and is used for further search and comparison of the profile in various social networks. At the current stage, the following data is used:

- First name, middle name of the user;
- Date of Birth;
- Place of residence;
- Place of Birth;
- Friends;
- Text notes (posts);
- Place of work;
- Place of study;
- Contacts, email, phone number;
- Profile avatar, as well as profile photos.

This information is downloaded, both for the original profile, and for the desired profiles in other social networks. The loaded profile data is mapped to the source profile data. Since there can be several profiles found, they are sorted according to the following criteria:

- Criteria for the presence of similar faces in photographs. Using the HOG method, people are found in photographs and their vector representation is generated. Subsequently, the Euclidean norms of the vectors are compared.
- The criterion for the presence of similar contacts. Profiles containing links to each other are very likely to belong to the same person.
- The criterion for a similar place of work and place of study. To calculate this indicator, the strings are pre-processed: they are cleared of punctuation, reduced to lower case. After that, the lines are lemmatized, and using the obtained lemmas, the lines are compared according to the following metric:

$$\alpha / (\max(\lambda 1, \lambda 2)), \quad (1)$$

where α is the number of pairwise matching lemmas, $\lambda 1$, $\lambda 2$ - the number of lemmas in lines 1 and 2, respectively. If the value of this criterion is more than 0.85, then the lines are considered similar.

- Criteria for the presence of similar posts. Two metrics were used to compare text notes. The first is finding the Levenshtein distance (editorial distance, editing distance)[12, 13] - the minimum number of operations to insert one character, delete one character and replace one character with another, necessary to turn one line into another. As a second method for finding similar posts, the shingles algorithm was implemented [14]. This algorithm works on the principle of splitting text into shingles, computing hashes of data of shingles, pairwise comparison of hashes. The following metric was used for the shingle method:

$$(2 * \nu) / (\sigma 1 + \sigma 2), \quad (1)$$

where ν is the number of matching hashs of shingles, $\sigma 1$, $\sigma 2$ - the number of shingles in the first and second row, respectively.

A visual representation of the shingle algorithm is shown in Figure 1.

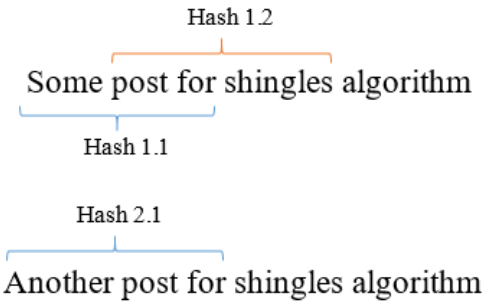

Fig. 1. Shingles Algorithm.

- Criteria for having similar friends. This indicator is calculated by pairwise comparison of the names of friends. The more matches, the higher the profile in the final search results.

IV. Implementation of a Software System for Mapping User Profiles on Social Networks

To test the effectiveness of the proposed approach, a software system for mapping user profiles on social networks was implemented. The developed system is a client-server application, where the server is a Java web service developed using the Spring Boot software platform.

The system integrates with the three most popular social networks in the CIS: Vkontakte [15], Odnoklassniki [16] and Facebook [17]. Data from the VKontakte social network is downloaded through the integration with VK API. Data from the Odnoklassniki and Facebook is downloaded by parsing the desktop and mobile versions of the website of the respective networks.

A web service containing an application in python has also been developed. This service, using the DLIB library [18], forms a vector representation for user photos.

As input, the system accepts a link to a profile in one of the social networks. From this profile, all possible data about a person is loaded, and according to these data a single model of the desired profile is formed. Similar profiles are searched by searching for fields loaded from the original profile. Based on loaded similar profiles, a rating is formed, according to which sorting subsequently occurs.

Similar profiles are sorted by the received rating and displayed on a web-form.

The system architecture is shown in Figure 2.

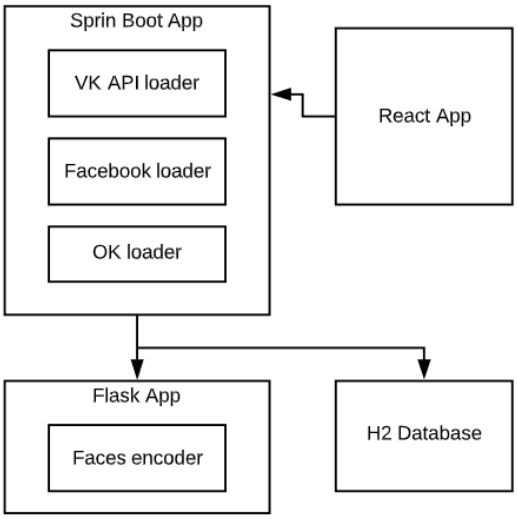

Fig. 2. System architecture.

As can be seen from the figure, the system consists of two server applications:

- Spring Boot App [19],
- Flask App [20]

There is also a React App client application. Spring Boot App contains the basic logic of the system, as well as data loaders from social networks:

- VK Api Loader,
- Facebook loader,
- OK loader

The Flask App contains methods for recognizing faces in photographs using the DLIB library. H2 Database is used to store some non-confidential data. An example of such data is the id of cities and countries from the VK API.

An example of the system interface is shown in Figure 3.

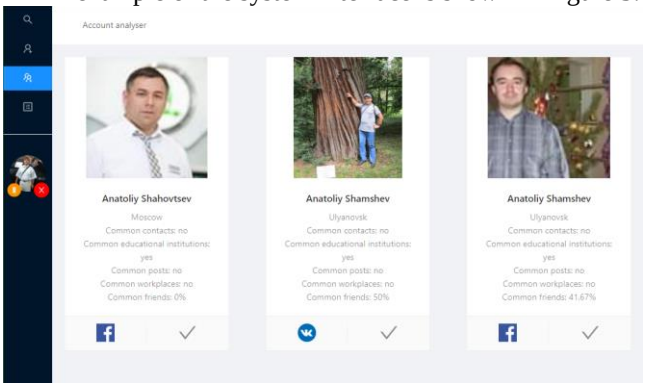

Fig. 3. System appearance.

The page displays the selected profiles, and on the left there is an application menu.

V. Experiment Results

A pre-prepared sample of 100 users with profiles in various social networks was used as an experimental base. All these users had 204 accounts, since not all of them had accounts in all networks at once. For each of these accounts, we tried to find similar ones using the developed service. As a result of the experiments, the diagram shown in Figure 4 was compiled.

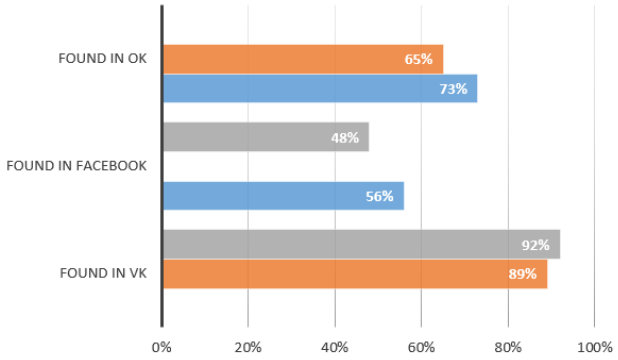

Fig. 4. Chart of the percentage of found profiles.

On the diagram you can see that the system coped best with finding profiles on the VK social network, and worst of all, Facebook. This is due to the convenience of extracting data from relevant resources. VK API allows you to quickly extract large amounts of data, which increases the quality of recognition, while parsing other networks consumes many resources, which forces to limit the amount of data retrieved.

The results for the profile comparison criteria were also calculated, the result is shown in Figure 5.

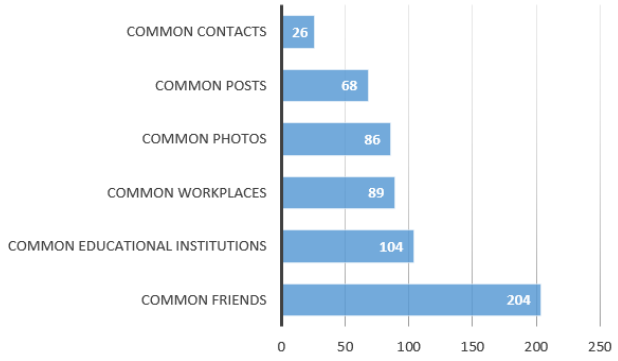

Fig. 5. Profile comparison criteria diagram.

On the diagram you can see that in all cases the system managed to find at least one common friend. The results of the coincidence of other criteria are much smaller. Twice worse, the system managed to find common educational institutions and common places of work. This is due to the fact that users do not always indicate these characteristics on their pages. Also, often the format of the specified data does not allow to correctly compare them. Even less, the system coped with finding common faces in photos. This is due to many factors, such as, for example, the accuracy of the model itself, the quality and number of photos uploaded. Only a third of the experiments managed to find common posts on the pages. This is due to the fact that users do not always fill pages with the same posts. Cross-references to profiles were found least of all, as users provide such information less often.

CONCLUSION

Thus, within the framework of this work, an integrated approach was proposed to find user profiles in different social networks by analyzing not only the data of profiles, but also poorly structured information from the pages of the respective accounts, as well as graphic information.

As a result of the work done, a software system was developed that performs the function of searching and mapping similar profiles on social networks. The application can be used as a personnel search platform. The proposed methodology lays the foundation for further work on conducting relevant experiments, developing new algorithms for searching, comparing, analyzing, and building a portrait of a user based on open data about they.

ACKNOWLEDGMENT

This work was supported by the Foundation for Assistance to the Development of Small Forms of Enterprises in the Scientific and Technical Sphere within the framework of the project "Development, technical implementation and testing of a prototype platform for the formation of a social portrait of an applicant based on intelligent data retrieval in social networks using the principles of knowledge engineering" of Agreement No. 60GS1CTS10-D5 / 56043 from 06.02.2020.

REFERENCES

[1] I. Rytsarev, D. Kirsh and A. Kupriyanov, “Clustering of media content from social networks using bigdata technology," Computer Optics, vol. 42, no. 5, pp. 921-927, 2018. DOI: 10.18287/2412-6179-2018-42-5-921-927.

[2] A. Filippov, V. Moshkin and N. Yarushkina, “Development of a Software for the Semantic Analysis of Social Media Content,” Recent Research in Control Engineering and Decision Making. Studies in Systems, Decision and Control, vol 199, pp. 421-432, 2019.

[3] N. Yarushkina, A. Filippov, V. Moshkin, G. Guskov and A. Romanov, “Intelligent Instrumentation for Opinion Mining in Social Media,” Proceedings of the II International Scientific and Practical Conference Fuzzy Technologies in the Industry, Ulyanovsk, Russia, pp. 50-55, 2018.

[4] Y. Gaewon, "Enhancing Entity Search with Social Network Matching," EDBT/ICDT: Proceedings of the 14th International Conference on Extending Database Technology, 2011.

[5] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," Proceedings of the eleventh international workshop on Web information and data management, 2009.

[6] E Raad, R. Chbeir and A. Dipanda, "User profile matching in social networks," 13th International Conference on Network-Based Information Systems, IEEE, 2010.

[7] J. Vosecky, D. Hong and V.Y. Shen, "User identification across multiple social networks," 1 international conference on networked digital technologies, IEEE, 2009.

[8] I. Veldman, “Matching Profiles from Social Network Sites,” Master's thesis, University of Twente, 2009.

[9] N. Yarushkina, A. Filippov, V. Moshkin, A. Namestnikov and G. Guskov, “The social portrait building of a social network user based on semi-structured data analysis,” CEUR Workshop Proceedings, vol. 2475, pp. 119-129, 2019.

[10] FindFace [Online]. URL: https://findface.pro.

[11] Yandex.People [Online]. URL: https://yandex.ru/people.

[12] V. Chernenkyi and Yu. Gapanyuk, “Passenger identification technique based on installation data,” Engineering Journal: Science and Innovation, vol. 3, no. 3, pp. 3, 2012.

[13] D.V. Mikhailov, A.P. Kozlov and G.M. Emelyanov, “An approach based on analysis of n-grams on links of words to extract the knowledge and relevant linguistic means on subject-oriented text sets,” Computer Optics, vol. 41, no. 3, pp. 461-471, 2017. DOI: 10.18287/2412-6179-2017-41-3-461-471.

[14] A.Tsimbalov and O. Zolotarev, “The method of shingles,” Vestnik of the russian new university. Series: Complex systems: models, analysis and management, vol. 4, no. 4, pp. 72-79, 2016.

[15] Vkontakte [Online]. URL: https://vk.com, last accessed 2020/05/11.

[16] Odnoklassniki [Online]. URL: https://ok.ru/.

[17] Facebook [Online]. URL: https://www.facebook.com.

[18] DLIB library [Online]. URL: http://dlib.net/.

[19] Spring Boot App [Online]. URL: https://spring.io/projects/spring-boot.

[20] Flask App [Online]. URL: https://flask.palletsprojects.com/en/1.1.x/quickstart/.