Iterative Search for Selecting Web-Pages to Construct Personal Profiles

Artur Harkovchuk, Dmitry Korzun
Petrozavodsk State University (PetrSU)
Petrozavodsk, Russia
{harkovch, dkorzun}@cs.karelia.ru

***Abstract*—Internet content provides much information about a particular person, especially related to her/his activity reflected in social networks as web-pages linked with information files. The problem is to find such information about a given person and with given search query constraints. A lot of data sources is had to be analyzed to find information for constructing digital personal profile. In this paper, we consider iterative Internet search algorithm that generates multiple search queries and evaluates the found links on the relation level to the given person. A rank is computed for each link that shows the relation level to the given person. Our experiments show that the method finds links among which 75% or more is essentially related to the person. When the content of a link is further analyzed (e.g., parsing) the search quality increases up to 90% and more.**

I. Introduction

Information is a key asset at the current time. Possession of relevant, processed data enables leadership in such areas as exchange trading or information data protection. Timely and full access to the published private information about a person allows you to immediately take measures to protect him.

With the development of such types of fraud both by telephone [1], by email, on social networks, it is necessary to apply measures for protection. One of these measures is to prevent the dissemination of your confidential information on the Internet. The methods of social engineering often use various hobbies and statuses of a victim. To ensure information security, it is necessary to know as soon as possible that private information(personal phone number, passport data, etc.) becomes publicly available on the Internet. This will allow to take measures to remove information as soon as possible.

The paper considers the developed method of automated search for data, which allows you to find information about a person published throughout the global Internet network. The developed algorithm allows you to automate the search for information on all indexed pages, which increases the amount of information received and reduces the time to receive it. The method used will allow each person to automatically get links to resources that contain information about him. This method will speed up the search for potentially dangerous content and reduce the man-hours required for this.

The rest of this short paper is organized as follows. Section II considers related methods for search in Internet content. Section III introduces the studies search problem. Section IV proposes our method with iterative query generation. Section V shows results of our early experiments. Section VI summarizes the key findings of this paper.

II. Related Work

This article presents a method for automating searching information about a person on the global Internet. The information found is used to build a personal profile. The application and methods of creating such profiles are described in [2] for the methods of solving socio-economic problems using digital technologies and Big Data. The following methods for analyzing a person are considered: identifying the interaction between people, building graphs of interests, evaluating the tonality of texts. These solutions are applicable for data that can be identified as information about the original person. In the present article, we consider a method for extracting such data for their subsequent analysis and construction of a personal profile.

Some identification methods can be found in [3]. They use ontology-based model for user profile building from information in social networks. This article describes identification based on lexicographic, lexical analysis, analysis of computer or network parameters. These methods can be used if the content is created by a person whose information you want to search for. The method of searching for information created by any person and posted on the Internet is presented in this article.

III. Information Search Problem for Digital Personal Profile Construction

Nowadays, every computer user constantly uses search engines. The search systems allow you to quickly get access to the information of interest.

For the best operation of a search system, a content is constantly scanned the Internet using robots to search for newly created web pages or updates. The next step is indexing. Robots attend the pages found during a scan to analyze them. After that, the link to the page will be available for search. All work carried out by search engines is aimed at providing the greatest coverage in the search, which makes them the best in the search.

At the current moment, the most popular search engine is Google. The Google search engine carries out more than 90 percent of search queries in the world. [4]. In different countries, this percentage may differ in connection with the country's policies. For example, in China, Google ranks only third with 6 percent. The first and second places are occupied by Sogou and Baidu with 60 and 25 percent, respectively. Therefore, it is worth considering the search engine used to search for information in a specific region. In our work, we will use the Google search engine, since it prevails over other search engines in Europe.

For a user, the search method is quite simple. It is enough to write an interesting phrase or question and the system will select the most likely references to resources. This method is not suitable for us, since the Google search engine, like other search engines, analyzes queries, searches for keywords, and searches. Sometimes it happens that search engines find what the user wanted, and not what he wrote.

The following Google search operators are used to build contextual search queries [5]:

- – Excludes words from search results;
- " " Finds accurate coincidence of the introduced phrase or word;
- OR Seeking content where one of the keys specified in the query is used;
- AND Displays the results of all keys (there may be 2 and more), which are combined with this operator;
- ( ) Groups multiple requests or operators.

Results of a search query are issued in a ranked order. Google uses over 200 ranking factors [6].

The Google search engine has several features that need to be taken into account when making search queries. The first feature is that if no results are found during a search query, the system notifies the user and offers results for a modified Google query. Before extracting links from the page, need to make sure that there is no such entry. If the record is present, then proceed to the next request. The second feature is that the names of classes in the web markup may differ depending on the browser and versions. It is necessary to unify the browsers used and the browser versions. The third feature is that Google prohibits automatic queries. It is required to apply tools that simulate the user's work to prevent blocking.

The links obtained as a result of the method allow building a profile of the person.

IV. Method with Iterative Query Generation

The developed method creates search queries and evaluates answers. The method works by compiling combinations of elements, evaluating connectivity and relationship, determining the cost of the request and compiling a query for a search engine, based on the syntax used by the search engine. Consider the stage of operation of the method related with finding the cost of each element in the query on the example shown in Figure 1. The figure shows the compilation of data sets and an example of the evaluation of each of the elements of the set (the evaluation of the element is in parentheses). The source data will be used: Name, Surname, city and nickname. For the method to work, the following parameters should be determined in advance: the cost of the data type, sets of related data and a numerical coefficient for each set, and sets of interacting elements with a cost for each set. The numerical values shown in the Figure 1 next to the sets are an example of cost estimation. Evaluation of the types of elements and sets is necessary for the method to work.

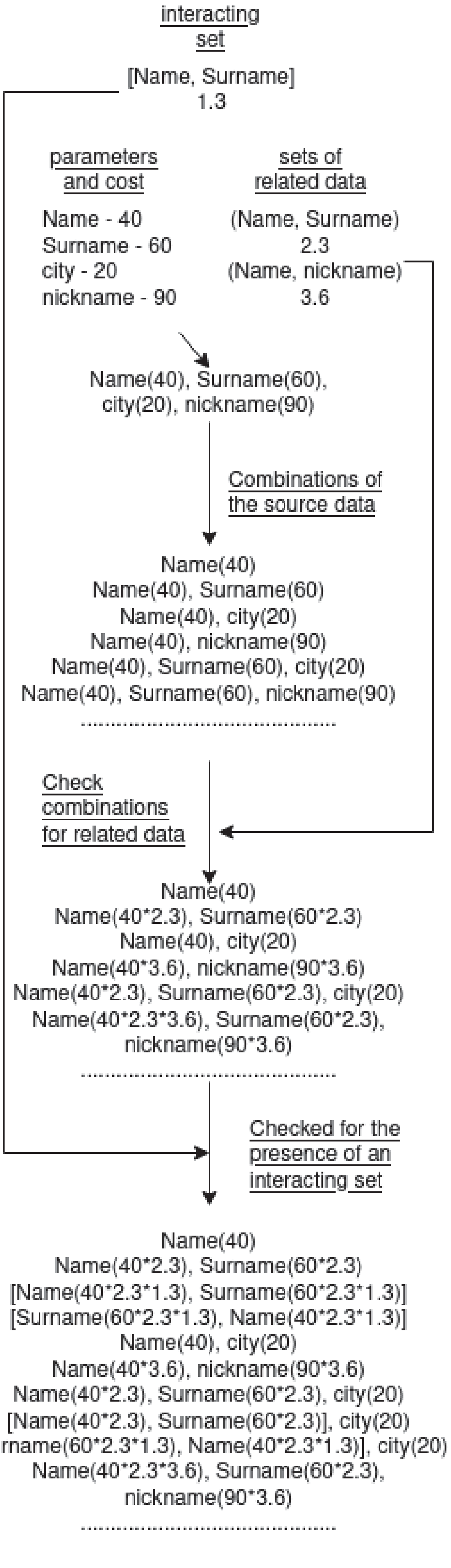

interacting set
[Name, Surname]
1.3
parameters and cost
Name - 40
Surname - 60
city - 20
nickname - 90
sets of related data
(Name, Surname)
2.3
(Name, nickname)
3.6
Name(40), Surname(60), city(20), nickname(90)
Combinations of the source data
Name(40)
Name(40), Surname(60)
Name(40), city(20)
Name(40), nickname(90)
Name(40), Surname(60), city(20)
Name(40), Surname(60), nickname(90)
Check combinations for related data
Name(40)
Name(40*2.3), Surname(60*2.3)
Name(40), city(20)
Name(40*3.6), nickname(90*3.6)
Name(40*2.3), Surname(60*2.3), city(20)
Name(40*2.3*3.6), Surname(60*2.3), nickname(90*3.6)
Checked for the presence of an interacting set
Name(40)
Name(40*2.3), Surname(60*2.3)
[Name(40*2.3*1.3), Surname(60*2.3*1.3)]
[Surname(60*2.3*1.3), Name(40*2.3*1.3)]
Name(40), city(20)
Name(40*3.6), nickname(90*3.6)
Name(40*2.3), Surname(60*2.3), city(20)
[Name(40*2.3), Surname(60*2.3)], city(20)
[Surname(60*2.3*1.3), Name(40*2.3*1.3)], city(20)
Name(40*2.3*3.6), Surname(60*2.3), nickname(90*3.6)

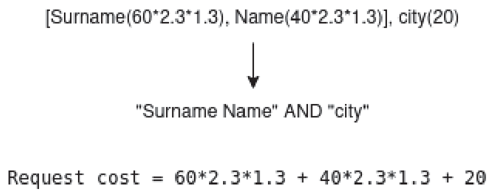
Fig. 1. Example for finding the cost of elements in a request

The cost of a data type is a numerical estimate of 1 to 300. This score is chosen empirically during experiments, depending on how the information identifies the person you are looking for. A set of related data is a set of information types that, when found in a single query, give a more accurate match with the person we are looking for information about. Each set is evaluated by a coefficient from 0 to 10.

A set of elements should contain together on a web page, is called as an interacting set. An example of such a set is the name and surname. The first and last names must be together so that the search results do not find information about people with the same last names, but with different names and vice versa. Such sets are also evaluated by the coefficient of 1 to 2.

The first step is to find all possible combinations of the source data. Each element in the combination is estimated by its cost which was empirically selected for each data type before the method began. The next step is to check each combination for the presence of elements in it that completely make up any of the existing sets of related data. If the corresponding data has been found, then the cost of each such element is multiplied by the cost of the found set of corresponding data. Next, the presence of an interesting set is checked. When found, 2 additional combinations are formed in which elements from the interacting set are combined. The first and second sets differ in the order of the elements. This is done because these two elements become one when searching and their order should be taken into account.

As a result of performing the stages, we get sets of elements with the cost of each of element from these sets. Figure 2 shows the example of converting the resulting set into queries and finding the cost of the query.



[Surname(60*2.3*1.3), Name(40*2.3*1.3)], city(20)

"Surname Name" AND "city"

```
Request cost = 60*2.3*1.3 + 40*2.3*1.3 + 20
```

Fig. 2. Example of creating a search query and rating

To compose a search query, each element is taken in double-quotes. The interacting set shown in Figure 2 in square brackets is also taken in double quotation marks, a space is placed between the elements. Put the word ”AND” between all the elements. The cost of the request is calculated by adding all the cost indicators of the elements, taking into account all the coefficients.

When building a search query, it should also be taken into account that some types have equivalent different values. For example, this can be a name. Anya and Anna are the same name, but if you search only for Anya, the results with the name Anna will be lost. Used () and OR to form a construction of the follow type - (”Anna” OR ”Anya”). The use of two different forms of names may be due, for example, to the fact that a person uses the first form, and in the news the second form is used.

Selenium is used to make the request. This tool allows simulating the browser as if it is done by a living person, which provide more time without blocking when working with the search engine. The links obtained as a result of the search must be ranked according to the following formula: ”Link cost = request cost - (request cost/ 10) + ((request cost/ 10) /Number of links taken from the request) * (Number of links taken from the request-the position of the current link + 1) ”. The cost of the link must be in a certain numerical interval. If the cost of the link is in the interval, it is considered that the page contains information about the person you are looking for. Using the formula allows the placement of a link in the search results to influence the final cost of the page found. Using the formula, we take into account the ranking of pages by the Google search engine.

V. Experimental Study

During the experimental study, the following results were obtained, as shown in Table I. The implementation of the method was written in Python. The data source is the social network Vkontakte. VK API was used to extract a data. Automatic queries were carried out using Selenium. The coefficients of elements and sets of elements were empirically selected.

TABLE I. Research of finding links about a person

ID in Vkontakte	Total number of pages found	The number of correctly found	The percentage of correctly found
id15438668	29	20	69%
id33514885	36	33	91%
id208445782	40	37	92%

During testing, it was revealed that the largest percentage of correctly found percentages are among people who have an active lifestyle: participate in competitions, engage in social activities, etc.x The experiment used the first 30 links from each search query. Using a large number of links from queries increases the number of pages found, but reduces the percentage of pages found correctly.

As an additional experiment, a facial recognition system was applied to photos of the pages posted on the found pages. If the system found matches, then such pages were considered to be correctly found. This method increased the accuracy of finding pages, but reduced their number.

It is based on the results that the developed method works well for finding information about a socially active person. If a person incorrectly indicates his name or surname, then it becomes almost impossible to find information about him. You need to know a lot of additional information to organize such searching for information that will allow you to find information about him.

VI. Conclusion

The article discusses a method that allows you to automate the search for information about a person on the Internet. Testing of the method showed that it can be used to perform an automated search for information only about socially active people. The results showed that this method can be used for automated search on indexed pages on the global Internet, to find information about a person of interest. For this method, the problem remains for persons who do not lead a socially active life. Information about such persons is impossible to find. In the future, it is necessary to develop an algorithm that allows you to automatically determine the coefficients of elements and change them in various situations.

The developed method can be also used to search person's second pages on various social networks. Testing of this method was carried out and showed a good result on finding pages that contain a reliable first and last name and side information.

ACKNOWLEDGMENT

This research is supported by RFBR (research project # 19-07-01027).

REFERENCES

[1] L. Peng and R. Lin, "Fraud phone calls analysis based on label propagation community detection algorithm," in *2018 IEEE World Congress on Services (SERVICES)*. IEEE, 2018, pp. 23–24.

[2] A. Y.Timonin, A. M. Bershadsky, A. S. Bozhday, and O. S. Koshevoy, "Social profiles — methods of solving socio-economic problems using digital technologies and big data," in *International Conference on Digital Transformation and Global Society (DTGS-2018), Volume 858 of the series Communications in Computer and Information Science (CCIS), Springer International Publishing Switzerland*, 2018, pp. 436–445.

[3] R. Niyazova, A. Aktayeva, and L. Davletkireeva, "An ontology based model for user profile building using social network," in *Proceedings of the 5th International Conference on Engineering and MIS*, 2019.

[4] "Top search engines in the world, statistics 2020," 2021. [Online]. Available: https://marketer.ua/search-engine-stat-world/

[5] "Google search operators: the most comprehensive list," 2021. [Online]. Available: https://seranking.ru/blog/seo/operatory-poiska-google/

[6] "200 google ranking factors in 2020: part 1," 2021. [Online]. Available: https://convertmonster.ru/blog/seo-blog/200-faktorov-ranzhirovaniya-v-google-v-2020-godu-polnyj-perechen-chast-1/