MATHEMATICAL MODELING AND COMPUTING, Vol. 5, No. 1, pp. 56–65 (2018)

Modeling Computing Mathematical

Ranking the social media platform user pages using Big Data

Mastykash O., Liubinskyi B., Topylko P., Penyak I.

Lviv Polytechnic National University,
12 S. Bandera Str., 79013, Lviv, Ukraine

(Received 9 June 2018)

The platforms of the social media of the Internet, depending on their content have been analyzed in the paper. The classification that allows selecting groups by specific one's signs has been made. To rank the pages of users of virtual communities, it is suggested to use a modified PageRank algorithm. An approach based on the use of lexical analysis and algorithm for ranking and organizing data using the MapReduce paradigm is developed. Using the developed approach and the appropriate algorithm, the software for ranking user pages has been implemented. The results of processed data and the formation of users' PageRank of the platform has been analyzed.

Keywords: *social media platform, big data, page ranking, measuring of page ranking, virtual community.*

2000 MSC: 90-04, 68-04, 68P10

UDC: 004.773.2, 004.45 **DOI:** 10.23939/mmc2018.01.056

1. The social media platforms work principles

The Internet social environment is nothing but what surrounds a user in his virtual life — that is a manifestation of relations between the users, which are caused by common interests, communication, ideas, activities, expression of opinions, content sharing, etc. [1, 2].

The Internet Social Environment Platform (ISEP) is a platform that provides the functional for the users communicating and allows you to integrate them into virtual groups (VHs). Virtual groups represent a group of users that are functionally interconnected within the platform — for example: community, friends, events, fan clubs, etc.

Users have an ability to communicate through the platform interface, which is unique to each platform of the Social Internet Environment. The generated content is stored in database platforms which work through special services that, in turn, implement a universal system for the data collecting and aggregating. Information in the databases is preserved mainly in the form of key-value pairs. ISEP consists of two parts:

1) Front-end part — provides the interface for the user's interaction with the platform services. This is the part which the user interacts with directly;
2) Back-end part — handles the user queries and serves the database access requests.

As a result we have the three-levels social media platform architecture:

1) Web Interface;
2) Data Processing Services;
3) Data Warehouse.

The main features of the Internet social media platforms are the tools for finding contacts and establishing the various types of relationship between their members, the ability to search the users. With the help of the virtual platform tools, every user can create his own virtual account — to consolidate a profile that expresses yourself. Considering such a kind of information the user's page will be able to find other members. The personal page presence already allows you to use the search mechanisms. The Internet Social Environment Platform delivers the following functionality [3]:

1) Creating and maintaining the personal page with the community member contact details;
2) List consolidation;
3) Create your own records and view the other user ones;
4) Communication between other participants;
5) Third-party services usage;
6) Ability to restrict the communication with the unwanted persons, etc.

The platforms are divided into the groups depending on the idea they are implementing. For the user's page data automated analysis the most important indicators are the platform content standardization, the user data access and the data representation architecture [4]. These indicators can vary dramatically on the different platforms of the same group of communities. Hence the following classification of ISEP is proposed according to:

- Data Availability Policies. When writing a content analysis system this feature affects the ability to receive page content by software that is not registered with ISEP or has no token to work with it:
    1) Content is fully available;
    2) Content is not available;
    3) Content is partially available;
- Page Stabilities. During the ISEP lifecycle its structure and architecture may vary:
    1) Staticly (it means the user page's structure does not change at all or changes insignificantly);
    2) Dynamically (the page structure changes periodically).
- User Definition. The completeness of the user profile description and the key nodes presence are the key features for the unique user identification:
    1) Information completeness of the user's description;
    2) Partial user description;
    3) Without any personal information.

2. Analysis stages

The process of the ISEP data representing along with its further analysis can be divided into the three stages (Fig. 1):

1) Data Receiving;
2) Data Filtering;
3) Data Structuring.

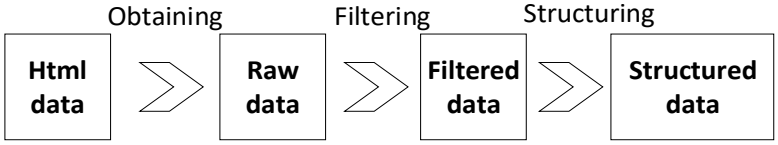

Fig. 1. Page analysis components.

The initial data analysis stage is the data obtaining along with their subsequent presentation [5]. The process of obtaining the data can be arbitrary: manual reading of data; ARI usage; the source of the page direct analysis; sending requests to the server; using automated tests, etc.

The data representation quality is one of the key parameters for the effective system functioning towards the social environment of the Internet platform analyzing. Depending on which platforms are analyzed and which key characteristics are needed the data representation process for the algorithm is different. You can submit data using a descriptive description of the object.

The quality of the algorithm work depends heavily on how much of the underlying input attributes really affect the result that is treated as the output. The more important and accurate the data is, the

greater is the likelihood the model would work better. Thus the presentation and design of an effective data analysis model begins with a selection of key features.

The next step in obtaining the data is their validation, that is checking whether the document was generated syntactically correctly. The following is the data filtering step. This stage is divided into the three components: Normalization; Filtration; Chopping.

The Normalization and Filtration processes are needed to remove the "information garbage", that is the portion of the information that does not have any benefit for the further analysis. The initial normalization document stage is it is verification for the correctness and the correctness of the encoding. The data is then given to the standardized form (register, word structure); there is a process of the "clearing the data" from the unnecessary tags during which only the page text remains. The data normalization stage allows it to be stabilized. In such way we get rid of the unnecessary information that would ensure better execution on the next steps. Already the filtration stage directly provides the removal of the unnecessary information (refer to the article or somehow describe the algorithm of filtration). The Chopping stage allows reducing the amount of the readable information by the repetitions removing.

Once the data is chopped and filtered it becomes possible to proceed with the next stage, namely to the data structuring (Fig. 2).

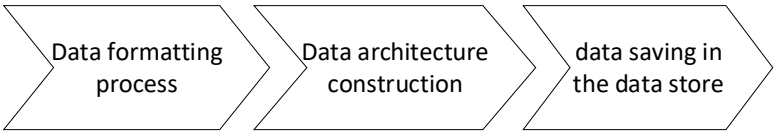

Fig. 2. Structuring data process.

This stage affects on how much quality and fast the software of the analysis of the page would work. The first step is the Data formatting process. Data from a text format should lead to another, more convenient for machine read format. Supported formats are the next: XML, JSON, Excel documents, etc.

The second stage is the Data architecture construction, that is the isolation of the types of nodes that are the subject of the further conservation and research. When scanning a user page you may get an unstructured large volume information [6]. To structure the received information it is necessary to develop a complex architecture of the underlying data storage. For example, if data is stored in the relational database then the resulting data must lead to the third normal form of the relational database (DB); if the data is stored in the JSON format then it is necessary to perform their serialization onto this format.

The third step is the data saving in the data store. To save the data you could use the one og the following databases: SQL Server, MySQL, SQLite, etc. When saving you can also use the respective ORM system, which will speed up the process of the writing DB queries.

3. Scanning and obtaining data algorithm

One of the main difficulties in developing a page analysis program is the users' behavior which is that about a half of all the social media platform users block their pages for unauthorized users [7–9]. To solve this problem, we need to run an application on behalf of the registered platform user. However, in the case of excessive activity, the program will be blocked by the server, so before the next page analysis you need to pause, which significantly reduces the operation speed. To speed up the work you can implement a number of challenging applications, where every thread refers to the server as a separate registered user and the list of pages is distributed among the threads.

This algorithm's work can be started on the several machines ("workers"). In such case the central server which acts as the manager should control the operation of all machines ("nodes"). All of them simultaneously analyze the individual user pages of the social network (Wi), generate the result and combine the partial (Ri) results into a single (Summary Result) and submit them to the database.

During the parallel stations work there is a high probability of errors, so the server should synchronize the work of individual stations and solve the appropriate conflicts.

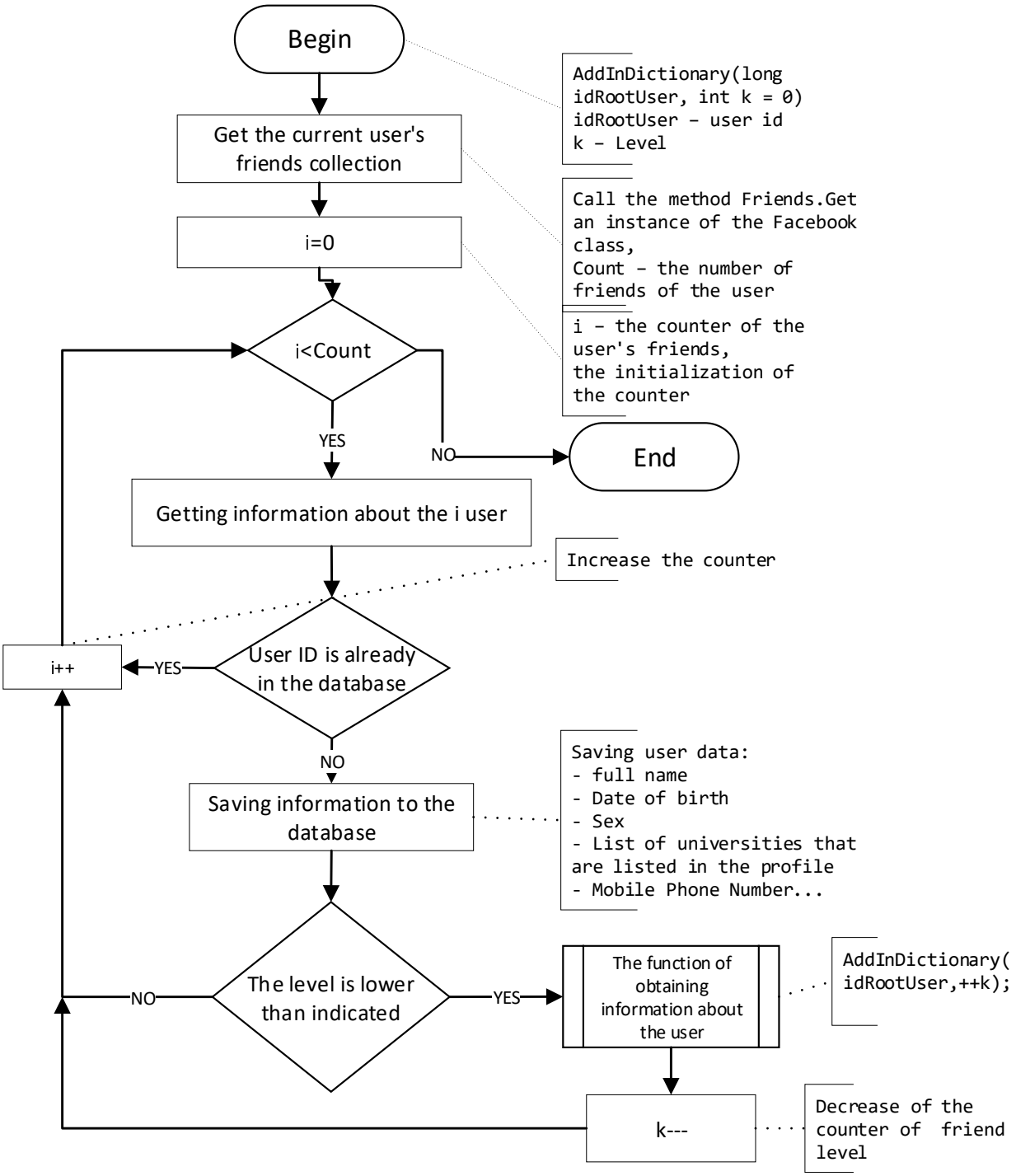

Fig. 3. The program algorithm.

The information obtaining from one platform user page is described in the article. To analyze multiple pages it is necessary to implement a recursive traversal of the platform pages.

During recursively browsing the network user's pages there is a high probability that one page will be analyzed and entered into the database several times. In order to avoid such situation, an

associative array with the identifiers of the users already passed might be created. If the user ID is already stored in the collection, is necessary to skip it.

The algorithm of the program is depicted in Fig. 3.

4. Platform user PageRank measuring

It is suggested to use the modified PageRank algorithm [10] to rank the pages of users of virtual communities. The basis of the algorithm is the following steps:

- Lexical analysis;
- Ranking algorithm;
- Ordering.

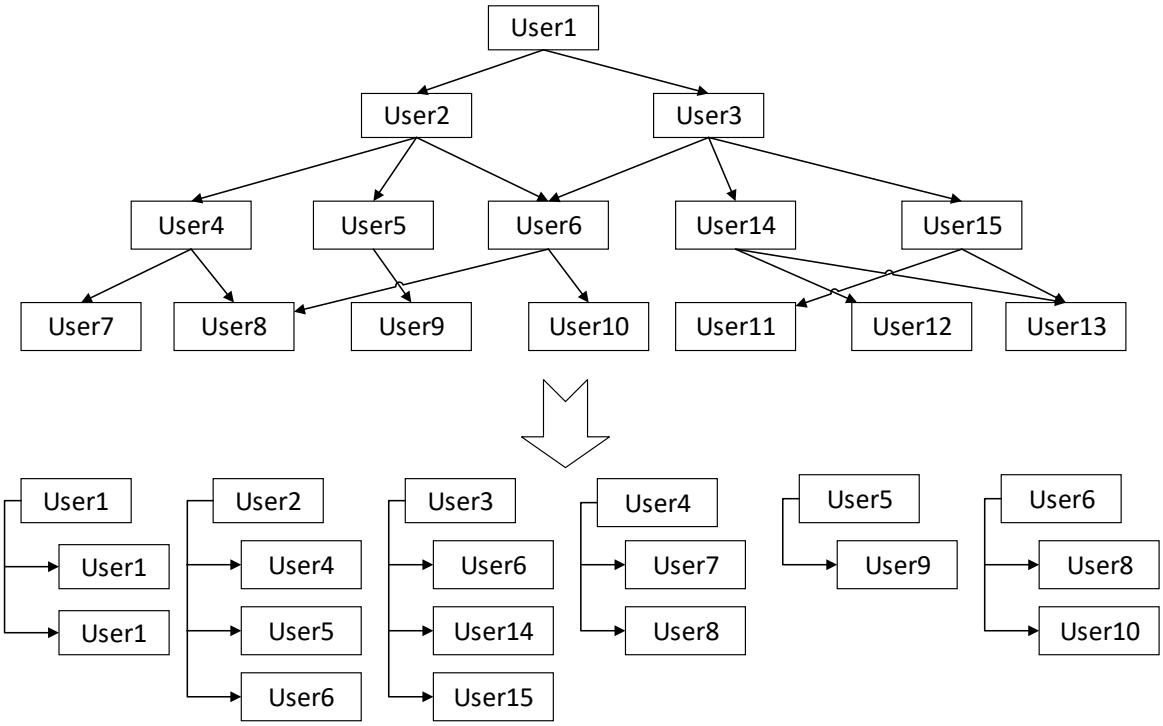

Fig. 4. Formation of the input file.

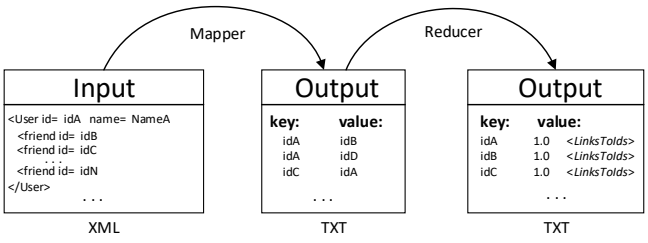

Fig. 5. Lexical data analysis stage.

At the stage of the Lexical analysis, the input data is treated as the XML document that was generated at the previous step (Fig. 4). The peculiarity of these documents is that their size is fairly large. The Lexical analysis is carried out in two stages. At the first stage, we get the formed document with the following structure — [key, value] where the key is the UserID and the value is the neighboring node identifier. At the second stage, the data is aggregated, analyzed and structured. The formed structure is similar to the structure obtained in the previous step, but may contain some differences: the key remains the UserID and the value is already represented as an object with two attributes where the first attribute is the initial PageRank and the second — a list

of identifiers for all the nodes to which it refers. Given a large amount of information for the Lexical analysis, it is suggested to use the MapReduce paradigm (Fig. 5).

At the ranking stage, the calculation of the new coefficient of importance of users of the platform of the Internet social environment. To implement the ranking process, it is also suggested to use the PageRank paradigm which is based on the two phases — mapping and reduce. In the mapping phase for the each input data structure a new structure is obtained where the keys are the UserID and the value is the neighbor's site identifier, the initial Page Rank and the number of outbound links. The reduce phase calculates the new value of the PageRank of the user's social media platform. The output data for the current phase is the data structured as follows: the key is the UserID and the value is the user's PageRank; the list of all the identifiers of the nodes to which it refers. The ranking stage is an iterative which repeats until the specified accuracy of the computations is achieved (Fig. 6).

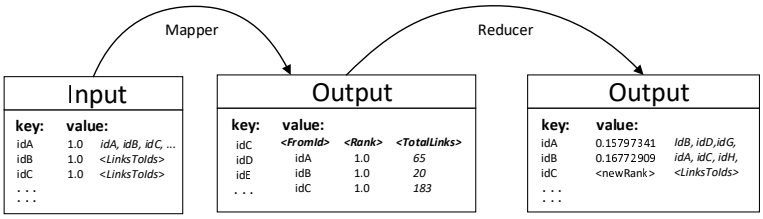

Fig. 6. Data ranging stage.

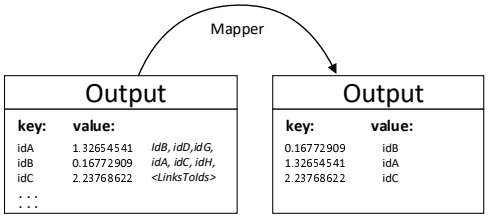

Fig. 7. Stage of data processing.

At the ordering stage, the second stage output is arranged into the data structure where the keys are the page rank of the user of the platform of the Internet social environment and the value of the UserID itself (Fig. 7).

5. Processed data analysis and the PageRank formation for the platform users

The analyzed social media Internet platform is quite large in scope (the number of users of some ISEP exceeds 100 million users). Therefore, a full ISEP scan and analysis may take more than one month. It was proposed to scan only a part of the whole platform rather than the complete one. For the partial scanning, a user (root node) is selected randomly in relation to which a recursive scan of all nodes is performed. To implement the data analysis, we introduce the concept of the level of the node proximity, which would mean how close the knot is resided towards the root. For example, a user with 0 as the near-level is a user who is his immediate friend, likewise a user with 1 as the level of closeness would be treated as a user who is associated with the previous user.

It is also proposed to perform a three-level analysis for the same node as well as for its related users with the level of proximity 1, 2 and 3 respectively. The user node is selected as the starting node having the 96 other nodes connections. Exploring the PageRank for the proximity 0 is an inappropriate, since all the neighboring nodes would have the single one root node reference and the root node will be referenced to all other nodes.

Table 1. Obtained results.

qUnique	qIter	time	maxR	minR
Proximity level1				
11920	12702	45 min	0.2002648115158081	0.15006278455257416
Proximity level2				
21154	22916	1 h 12 min	0.21385052800178528	0.15004640817642212
Proximity level3				
5 949 829	6 247 716	10 h 51 min	0.22514712572097778	0.15004640817642210

The program implementation results for the three-level system of nodes closeness are represented in Table 1, where qUnique is the number of unique users received, qIstr is the number of iterations

passed, time is the total time of the program run, minR is Minimum PageRank, maxR is Maximum PageRank.

The results of the user PageRank for each level of proximity are visually presented in charts 8–10 and in Tables 2–4.

Table 2. Proximity Level1.

Page Rank	Level1 user amount	Level2 user amount
0.150062784	0	596
0.150062784	0	596
0.150062784	0	596
0.150104805	0	596
0.150104805	0	596
0.150118470	0	596
0.150163918	0	596
0.150163918	0	596
0.150179922	0	596
0.150189712	0	596
0.150196060	0	596
0.150242701	0	596
0.150272741	0	596
0.150281056	0	596
0.150285407	0	596
0.150329098	0	596
0.150405690	0	596
0.150530889	0	596
0.150895878	0	596
0.200264811	96	500

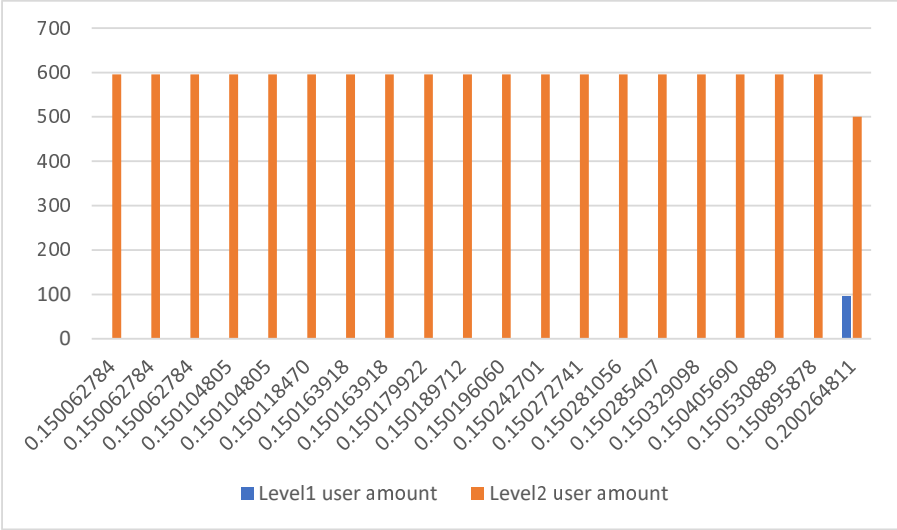

Fig. 8. Proximity Level1 histogram.

After analyzing the results, it could be argued that the greater the number of levels of the proximity is, the more accurate the research results are to be expected.

Table 3. Proximity Level2.

PageRank	UserId	Level1 user amount	Level2 user amount	Level3 user amount
0.150046408	192425425	0	0	1057
0.150046408	180636404	0	0	1057
0.150068074	125611634	0	0	1057
0.150069743	224406695	0	0	1057
0.150069743	183661323	0	0	1057
0.150070235	18326784	0	0	1057
0.150083005	154454580	0	0	1057
0.150083005	225669201	0	0	1057
0.150168418	5287940	0	0	1057
0.150257170	4685867	0	0	1057
0.150286450	21264540	0	0	1057
0.150304093	12870387	0	0	1057
0.150338605	17970164	0	0	1057
0.150369822	122839182	0	0	1057
0.150468364	49902159	0	0	1057
0.150507017	5080647	0	0	1057
0.150590553	152067995	0	0	1057
0.150833308	93319707	0	0	1057
0.151060581	65863383	0	81	976
0.213850528	4081433	96	103	858

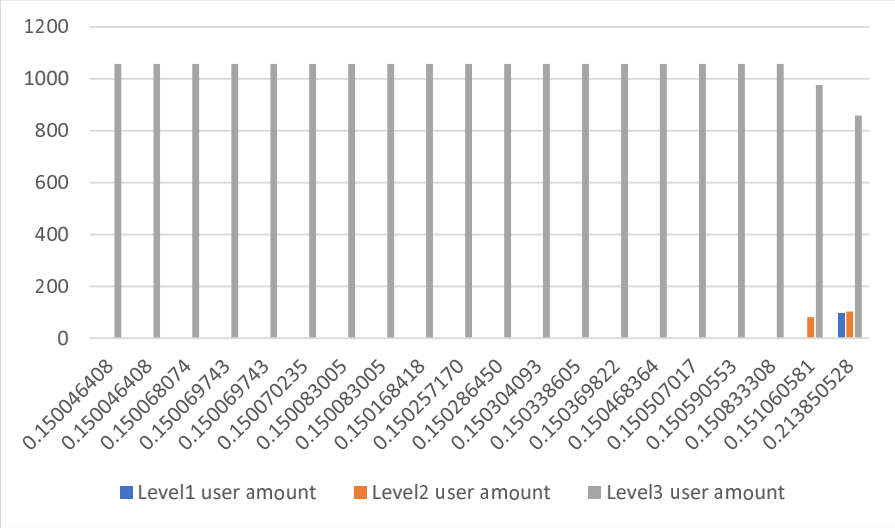

Fig. 9. Proximity Level2 histogram.

Table 4. Proximity Level3.

PageRank	UserId	Level1 user amount	Level2 user amount	Level3 user amount	Level4 user amount
0.150046408	192425425	0	0	0	297491
0.150046408	180636404	0	0	0	297491
0.150068074	125611634	0	0	0	297491
0.150069743	224406695	0	0	0	297491
0.150069743	183661323	0	0	0	297491
0.150070235	18326784	0	0	0	297491
0.150083005	154454580	0	0	0	297491
0.150168418	225669201	0	0	0	297491
0.150289028	5287940	0	0	0	297491
0.150331601	4685867	0	0	0	297491
0.150355488	21264540	0	0	0	297491
0.150416657	12870387	0	0	0	297491
0.150468364	17970164	0	152	20441	276898
0.150480642	122839182	0	1021	25811	270659
0.150468364	49902159	81	2418	40827	255640
0.150649055	5080647	256	3209	45817	248209
0.150833308	152067995	438	5042	51934	240077
0.150885045	93319707	831	7055	60634	228971
0.155854776	65863383	964	8534	67339	220654
0.225147125	4081433	1488	11357	85856	198790

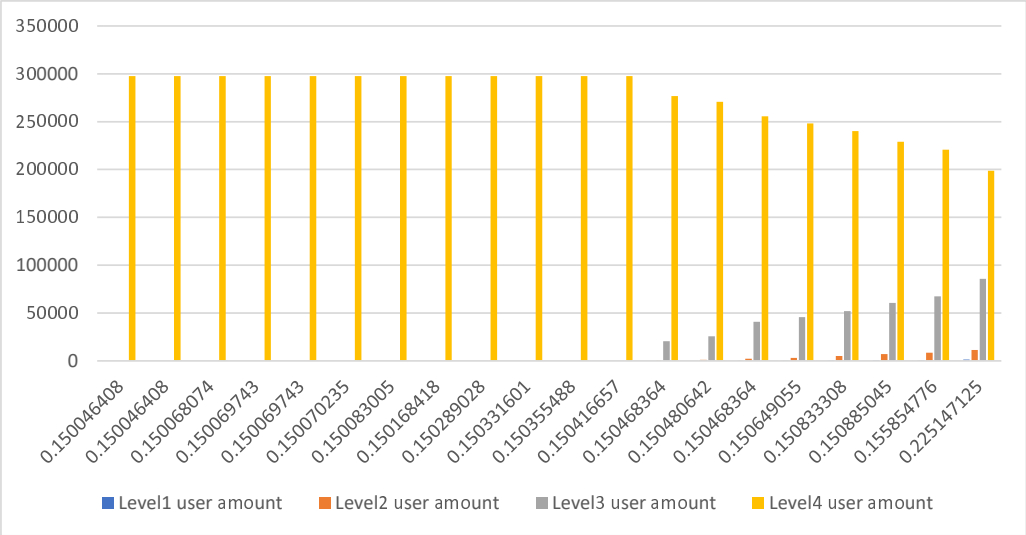

Fig. 10. Proximity Level3 histogram.

6. Conclusions

The paper analyzes the basic work principles of the Internet social media and proposes the classification of the platforms of the Internet social media; in particular the following characteristics are the basis for such division: data availability policy, page stability, completeness of the user's description.

The main stages of the ISEP user page analysis were selected: data acquisition, data filtering, data structuring. At the data acquisition stage, the characteristics that fall within the analysis and are implemented by the software that reads the data from the ISEP are indicated. The second stage is the data architecture design, that is the isolation of the node types of that are subject for the further conservation and research. At this stage the data from an unstructured look is brought to structure. The third step is to store data in the data store.

The approach of finding the importance of the ISEP users is proposed which is based on the stages of the lexical analysis, ranking and the data ordering by means of Big Data.

[1] El Morr C., Maret P. Virtual Community Building and the Information Society: Current and Future Directions", IGI Global (2012).

[2] Trach O., Fedushko S. Development of Software Complex of Virtual Community Life Cycle Organization. International Journal of Computer Science and Business Informatics. **17** (1), 1–11 (2017).

[3] Trach O., Peleshchyshyn A. Development of directions tasks indicators of virtual community life cycle organization. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017. 127–130 (2017).

[4] Mastykash O., Peleshchyshyn A. Analysis of the Methods of Data Collection on Social Networks. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017. 175–178 (2017).

[5] Shakhovska N., Vovk O., Hasko R., Kryvenchuk Y. The Method of Big Data Processing for Distance Educational System. In: Shakhovska N., Stepashko V. (eds) Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing. **689**, 461–473 (2018).

[6] HowardṪ. Design to Thrive: Creating Social Networks and Online Communities that Last. Elsevier (2015).

[7] Luo Q., Cui H., Zhang B., Zhang D. Ranking social network objects. US Patent 9,081,823 (2015).

[8] Papadopoulos S., Kompatsiaris Y. Social multimedia crawling for mining and search. Computer. **47** (5), 84–87 (2014).

[9] Zuckerberg M., Sittig A. Mapping relationships between members in a social network. US Patent 9,183,599 (2015).

[10] Lu X., Liang F., Wang B., Zha L., Xu Z. DataMPI: extending MPI to hadoop-like big data computing. In: Parallel and Distributed Processing Symposium, 2014 IEEE 28th International. IEEE, 829–838 (2014).

Ранжування сторінок користувачів платформ соціальних середовищ Інтернету засобами Big Data

Мастикаш О., Любінський Б., Топилко П., Пеняк І.

Національний університет «Львівська політехніка»,
вул. С. Бандери, 12, Львів, 79013

Проаналізовано платформи соціальних середовищ Інтернету залежно від їхнього контенту. Здійснено класифікацію, яка дала змогу виокремити групи за певними ознаками. Для ранжування сторінок користувачів віртуальних спільнот запропоновано використовувати модифікований алгоритм PageRank. Побудовано підхід, який основується на використанні лексичного аналізу, алгоритму ранжування та упорядкування даних з використанням парадигми MapReduce. Реалізовано програмне забезпечення для ранжування сторінок користувачів. Проаналізовано результати оброблених даних та формування PageRank користувачів платформи.

Ключові слова: *соціальна медіа-платформа, великі дані, рейтинг сторінки, оцінка рейтингу сторінок, віртуальна спільнота.*

2000 MSC: 90-04, 68-04, 68P10

УДК: 004.773.2, 004.45