Identification and Modeling of Historiographic Data in the Content of Web Forums

Nataliia Khymytsia[1][0000-0003-4076-3830], Taras Ustyianovych[2][0000-0002-6323-7924], and Ivanna Dronyuk[3][0000-0003-1667-2584]

Lviv Polytechnic National University, Lviv 79013, Ukraine
nhymytsa@gmail.com[1], ustyk5@gmail.com[2], ivanna.m.droniuk@lpnu.ua[3]

Abstract. The research has developed a series of steps to detect and simulate historiographic information on web forums through Web-Scraping, Data mining and Big Data analytics. The system of posting ranking is described in order to determine the relevance of the content for the needs of the historian, developed an algorithm that determines the significance of the system, its impact on the quality of the historical research. The given code snippets in the SQL query language to get the most useful aggregate numeric and text data will help simplify the research methods and develop new ways to identify knowledge and information in the content of web-forums of historical topics. Monitoring of web-forum user's activity using time series analysis has been completed. The rush hour for the generation of historiographic data in selected thematic sections of the web forum was determined. It is substantiated that processing technologies use for a large number of text data using Natural Language Processing (NLP) and Deep Learning will also allow automating the detection of valuable characteristics and features of the obtained data.

Keywords: Social Networks, Historiographic Information, Python, User Behavior, Timeseries Data, Web Scraping, Data Mining, Data Warehouse.

1 Introduction

Informatization has led to an exponential increase in the amount of information, the creation of local and global systems and networks, databases and knowledge, the emergence of fundamentally new technologies that radically changed the methodology of historical research. The intellectual and informational potential has become the result of informatization, and the development of both depends on the intensity of the information society process. These two concepts can be combined in one - the information and cognitive potential that characterizes the process of informatization. An important component of information and cognitive potential is the intellectual potential, which manifests the ability of a person to solve problems using the accumulated knowledge, skills and experience. The second component is the information potential, which provides the necessary level of awareness of members of society, that is, the ability to summarize, search, store and transmit information. Accumulation of information and intellectual potential as a result of informatization is the accumulation of

historiographical information and the formation of a new resource of historical knowledge in the Internet environment.

Exponential growth of electronic resources, observed at the beginning of the XXI century, opens new perspectives for the development of historical research. The global computer network of the Internet is a complete source of professional information for historians. An additional tool for cognition for historians is the study of a large array of unstructured historical information generated in web forums. To do this in Digital history, it is important to use various methods of data mining to automatically detect web documents, retrieve information from web forums, and identify its general patterns on the Internet.

2 Related Work

Scholarly works include a variety of areas for applying knowledge and data consolidation, using a wide range of resources for research on web communities. E. Trunzer examines high-performance architecture for gathering and consolidating data from multiple sources and resources into a single repository, their preparation for processing, and Big Data Analytics [1]. The web forum content usage and social network data analysis for conducting qualitative research, as well as the processes for data collection, and new knowledge and information acquisition from their processing is considered in B. McKenna, M. D. Myers and M. Newman.

The indicated methods of carrying out qualitative research of data of information systems, ways of preparation and data collection with the help of Web scraping frameworks with the purpose of obtaining unique scientific results by researchers in various fields of science [2].

The transformation of large amounts of data into Smart Data in a wide range of digital humanities (Digital History, Digital Sociology, etc.) is defined in the work of M. Zeng, which describes ways of transforming "raw" data into useful for historians and humanities information in order to identify additional knowledge [3].

In N. Khymytsia, T. Ustiyanovich work the importance of the research is that it presents Big Data and history usage [4]. Scientists P. Zhezhnych, N. Khymytsia, S. Lisina, O. Morushko in their study carried out a comprehensive analysis and systematization of mathematical and computer-oriented methods of processing historical information and generalized the historical experience of using information technologies in the Ukrainian historiography of science [5]. In a study by K. Artem, N. Kunanets, R. Holoshchuk, V. Pasichnik and A. Rzheuskyi conducting scientific research on the electronic science platform requires the establishment of effective communication between virtual team members [6].

At the same time, in the context of our study, valuable works are those that deal with virtual communities flow. It is examined in K. Miller studies [7]. Web communication formation and text mining refers to H. Rheingold's works [8]. Other authors consider virtual communities as a means of communication and education. In particular, N. Kristakys and J. Fowler on the analysis of social networks show that network

activity is productive [9]. Works by the scientists R. G. Howard are valuable for our study as well [10] and Yalan, Y., Xianjin, Z., Jinchao, Z., & Xiaorong, H, [11].
Historiographic aspect of information about the situation in the ATO area in virtual communities is examined in the works of A. Peleshchyshyn and N. Khymytsia [12, 13]. Language and socio-demographic differences of Internet communications are explained and covered in scientific publications of S. Fedushko [14]. Communication interaction features based on Web forums are analyzed in details by O. Tymovchak-Maksymets, O. Trach, V. Vus [15]. In the works of T. Bilushchak, A. Peleshchyshyn, M. Komova the historical information searching technology, taking into account information potential, means of observation, and a retrospective analysis of events development were considered. The algorithms of search and identification of Internet sources of historical facts and preconditions that influence truth of historical event witnesses or the author Internet source are presented [16]. System for text data analysis and mining from web-forum content, which helps to determine author contibution to specific text work or web-post, was developed and described in the work of I. Khomytska, V. Teslyuk, A. Holovatyy, and O. Morushko and others [17].
Spam and discussion detection using NLP tools and text mining methods is developed in the research of Y. Chen and H. Chen. Scraping of web-forum based data help to collect huge amounts information and get it processed [18]. Linguistic method for web-content comparison is used in the scientific work of P. Zhezhnych and O. Markiv, using documentation data and automated tools for automated filling of tourism documentation via web-forum content [19].
The purpose of our study is to apply interpreted object-oriented programming language Python for data collection, integration into the database; use of the data warehouse and the language of SQL queries for processing the historiographic information of web forums and determining the relevance of the data obtained for historians.

3 Main Part

In modern conditions, changing the paradigm of historical science, the use of various research resources of the Internet is a particularly urgent task. Within the framework of Digital history, scientists will actively involve various information resources for the study of recent history; sites that cover virtual and real scientific and historical communication; sites that have sources and special research on history; sites of scientific and social funds that support historical scientific and educational projects. Among the wide variety of modern research resources of the Internet, it is worth highlighting the source-research potential of direct-communication resources, since they maximize the function of communication, and therefore generate the primary and secondary sources of historical information. The specifics of directly-communicating Internet resources (partnerships, forums, social networks, blogs, multimedia web communities, and so on) lies in their in interactivity, embodied in the very technology of the WWW.
The most popular types of web communities are forums that are designed to communicate with network users. Features of social network historical information

are these: information content is unstructured, discussions arise spontaneously around various information drives (photos, posts of participants, discussions) [20].
A historian who uses web forum content should keep in mind that information for various reasons may be removed or temporarily closed to third-party users. Therefore, the researcher faces the task of quick identifying, collecting and processing thematic information. For qualitative analysis of unstructured and voluminous historiographic information, it is important for historic experts to use the best methods of intelligent analysis, based on Web Mining system and Data Mining technology.
Data processing involves several stages that will enable the collected data to be consolidated and integrated into warehouse, make it accessible for later processing, provide historians with convenient and reliable work with the information they receive. Thus, in Fig.1, the main components of the data workflow from the web forums into historiographic information database are highlighted.

Fig. 1. Data workflow blockscheme.

The first step involves generating a request for historical information source (web forum), parsing data with R/Python programming languages libraries and functions according to website structure (see Fig.2). Each structural element has its own tag,

which is fairly easy to obtain by parsing and web scraping techniques, but one of the major drawbacks of this method is the periodic change of class names and identifiers contained within the HTML tag structure of the site. The frequent parsing problem of web forums is data encoding (Javascript tags within HTML-tags), which requires additional requests (Ajax request etc.) generation and data decoding techniques.

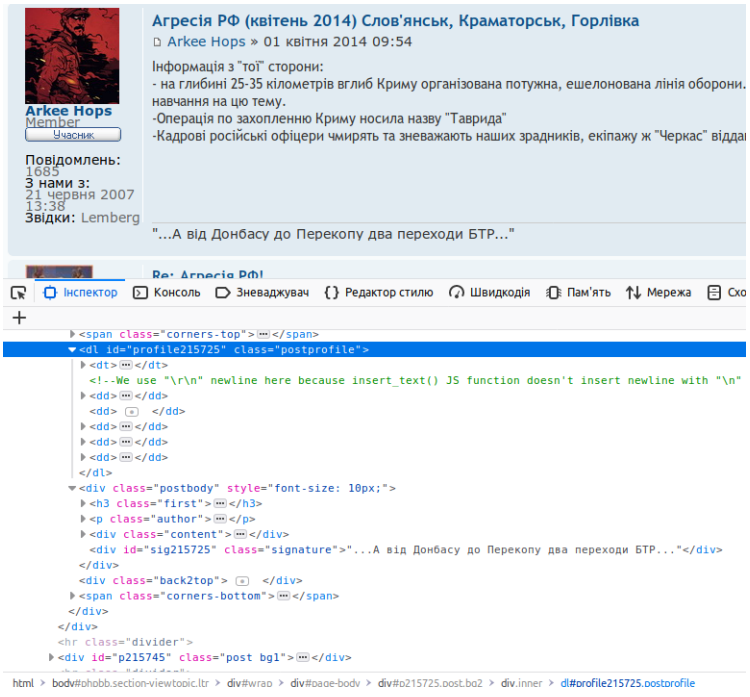

Fig. 2. Historical web-forum HTML structure.

After that, data characteristics with applying of natural language processing/ understanding (NLP, NLU) are determined automatally, which helps to obtain the maximum amount of useful information, such as: keywords, presence of discussion in the post, negative or positive post content and other features. So, the collected data must be integrated into data warehouse or relational database (see Fig.3).

The main table - "Posts", consisting of 10 characteristics and has relations to other tables. Automated data processing and the ability to use regular expressions, NLP / NLU and Deep learning will allow us to identify such characteristics as: a keyword, a historical events that are described in the posts, the presence of discussion in the post, chronological boundaries, topics, etc.

Extracted and integrated into the database posts will be sorted by using the described below data arranging method, which allows to highlight the most relevant and interesting posts and those that contain useful for historians information.

The next step is to determine the effectiveness of the ranking system by comparing the ID order by increasing of the unique post ID (Post_ID in the main table of the "Posts" database), which is the initial state; and after the post ranking process and

their new sorting order definition in the main table. In order to execute this, an algorithm has been developed that identifies the difference between two sets of consolidated data.

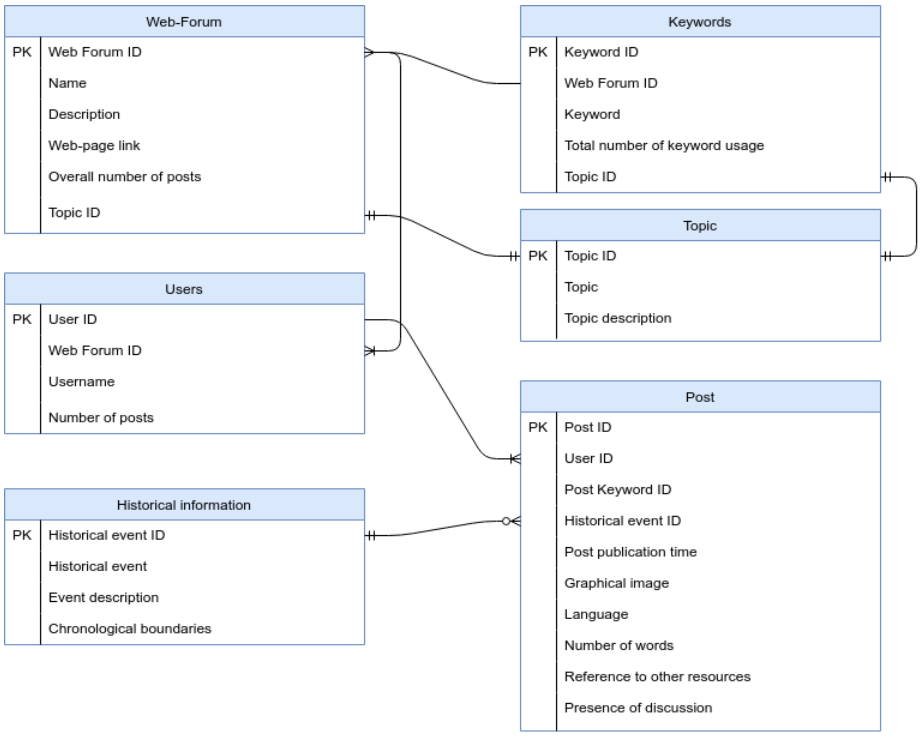

Fig. 3. Database structure and design for historical data storing.

The algorithm works as follows: two datasets are accepted at the input, which must contain the post identifier (Post_ID). The first set of data - the initial one, which was not used for post ranking, the second set - the one that "passed" the post ranking system. After that, the input table length is determined. If it is equal one to another, depending on the position of the posts and their identifiers in both sets of data, the similarity between argument 1 and argument 2 is calculated. The result is percentage of similarity. The lower the percentage of similarity is, the greater significance the post ranking system has. If the dataset length differs, the program does not perform computations.

```
# Python 3.6
def calculate_difference(dataset1, dataset2):
    count = 0
    dim = len(a)-1
    if len(a)==len(b):
        for i in range(0, len(a)-1):
            if a[i]==b[i]:
```

```
                count+=1
            elif a[i]==b[i-1] or a[i]==b[i+1]:
                count+=0.9
            elif a[i]==b[i-2] or a[i]==b[i+2]:
                count+=0.8
            else:
                count+=0
        diff = count/dim * 100
        return "The similarity between 2 methods is
{}%".format(round(diff,2))
    else:
        return "Different length of input datasets"
```

The final stage is ready-for-processing data extraction for historians, which will greatly simplify their work, and allow a qualitative summarizing and hypothesis development and building based on data analysis of posts from web forums containing historiographical information. Data querying from an existing warehouse is possible by using the SQL query language, working with relational databases. Complex queries formation for obtaining valuable information will allow getting the most necessary information in a short time spans and without high resource usage.

```
--SQL Query to count number of posts per web-forum with filter-
ing;
SELECT COUNT(P.POST_ID), U.WEB_FORUM_ID FROM POSTS P
LEFT OUTER JOIN USERS U ON U.USER_ID=P.USER_ID
WHERE LANGUAGE=<STRING> AND PUBLICATION_DATE > <DATETIME>
GROUP BY U.WEB_FORUM_ID;
--Query to intersect keywords of various topics;
SELECT KEY_WORD FROM KEYWORDS WHERE TOPIC_ID = <ID>
INTERSECT
SELECT KEY_WORD FROM KEYWORDS WHERE TOPIC ID != <ID> OR TOPIC_ID
< <ID>;
--Query for summing up all user's number of posts grouped by
web-forum
SELECT WF.WEB_FORUM_NAME, SUM(U.POSTS_NUM) from USERS U INNER
JOIN WEB_FORUMS WF ON U.WEB_FORUM_ID = WF.ID
GROUP BY WF.WEB_FORUM_NAME
HAVING SUM(U.POSTS_NUM) >= <NUMBER>;
--Add filtering (optional)
--T-SQL stored procedure to compute difference between number of
discussions in posts with different languages;
CREATE PROCEDURE Compute_Difference
  (
  @Crt1 VARCHAR(20),
  @Crt2 VARCHAR(20)
  )
```

```
AS
BEGIN
  DECLARE @Dataset1 INT
  DECLARE @Dataset2 INT
  DECLARE @Diff INT

  SET @Dataset1 = (SELECT COUNT(ID) FROM POSTS WHERE
DISCUSSION=TRUE AND POST_LANG = @Crt1)
  SET @Dataset2 = (SELECT COUNT(ID) FROM POSTS WHERE
DISCUSSION=TRUE AND POST_LANG = @Crt2)

  SET @Diff = ABS(@Dataset1 - @Dataset2)
  PRINT @Diff
END
```

In addition, the use of ranking system will allow us to receive up-to-date information quickly. Historiographic post ranking methods: summing up the existing data characteristics, which will allow to calculate the information content of a separate post stored in the database. For the posts ranking system, the following characteristics will be used from the "Posts" table of the historiographical web-forum information database: post date and time publication, number of keywords, presence of graphic images / illustrations in the post, appropriateness of current number of words with the optimal number of words, reference to certain resources, and presence of discussion. Therefore, most features will be sorted by the system and / or filtered by the end user (a specialist, historian). The appropriateness to the optimal length of a post calculation is possible using the following equation:

$$V_i = |N_i - N_o|, \tag{1}$$

where V_i – value that defines the optimal post's length. The lower this value, the higher is the appropriateness;
N_i – number of words in the n_i post;
N_o – the optimal number of words for a web-forum post.
Characteristics that correspond to the logical values True or False will respectively represent the value displayed on the equation below:

$$a_i = \begin{cases} a_i = 1 & a_i = True \\ a_i = 0 & a_i = False, \end{cases} \tag{2},$$

where a_i – a characteristic that corresponds to the value True or False (the presence of images in the post, calls to a certain resource, the presence of discussion in a post etc.)
User activity monitoring in a web-forum is an important feature that should be taken into account during the stage of conducting exploratory data analysis. It will reveal certain patterns of online activity, execute some kind of timeseries analysis and obtain information about when most of data in the historical web forum content is generated. Powerful libraries and tools that are part of Python's programming language allow us to collect huge amounts of data from the Internet. For timeseries analysis and activ-

ity monitoring, these tools were used, and the transformation & visualization of the collected data was carried out in order to obtain certain insights on forum users activity. Timeseries analysis is a systematic approach by which mathematical and statistical questions are answered posed by time correlations [21].
Most research information products and ways to improve them occupies precisely analysis of user behavior, in particular their activity [22]. On Fig. 4. reflects the activity of users of the historical web forum in a separate thematic block, discussion of historical issues which lasted for a month.

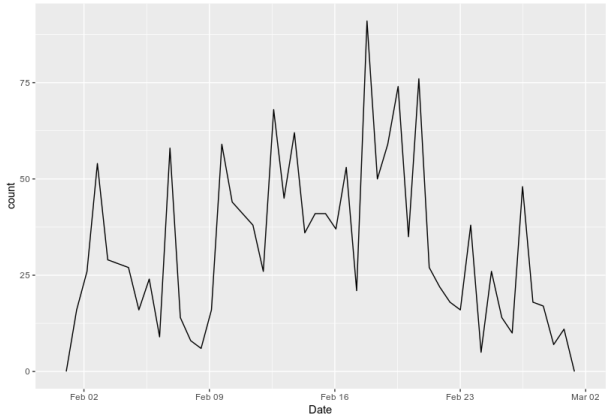

Fig. 4. User activity over the time of the existence of the topic of the web forum.

From this graph, we can see that activity was almost always the same and quite stable (about 50-75 posts daily), the peak of activity is at the beginning of the second half of the month. That is, the most active discussion in the forum occurs when the number of posts crosses the middle, after which the activity lasts for a certain period of time and gradually decreases.

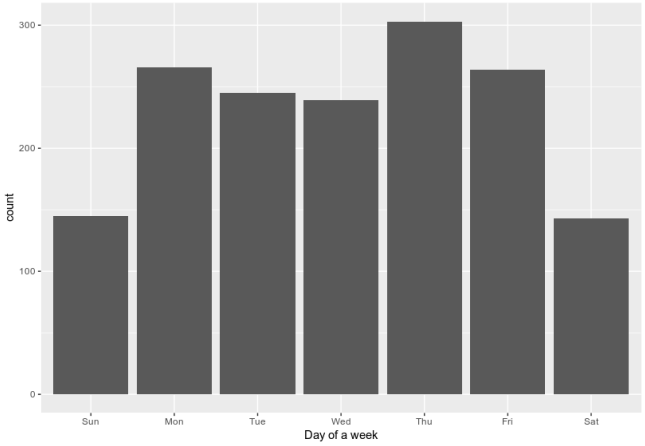

Fig. 5. User activity by day of the week.

The key point is to determine the weekday of deploying a discussion on web forums, especially relevant for historical topics, which allows you to determine when users themselves are able to participate in resolving disputed issues of the past as well as the present. So, as we can see in the graph, most users of historical, military-historical web forums are active on weekdays, whereas on weekends they are less and less time-consuming. High activity is observed at the beginning of the week, but gradually it decreases. Later it rises again (Thursday), and then it finally becomes minimal (Saturday, Sunday).

It's important to watch the time of the day when new content is being published on a web forum. This allows the user, including the historian, to choose the best time to publish his own post so that other members of the web forum can familiarize with and respond to it. At the web-site under investigation, the peak of activity was at lunch (13: 00-15: 00). If we take a gap from 5:00 to 20:00, we can observe a normal distribution, while from 00:00 to 5:00 - the distribution of Poisson.

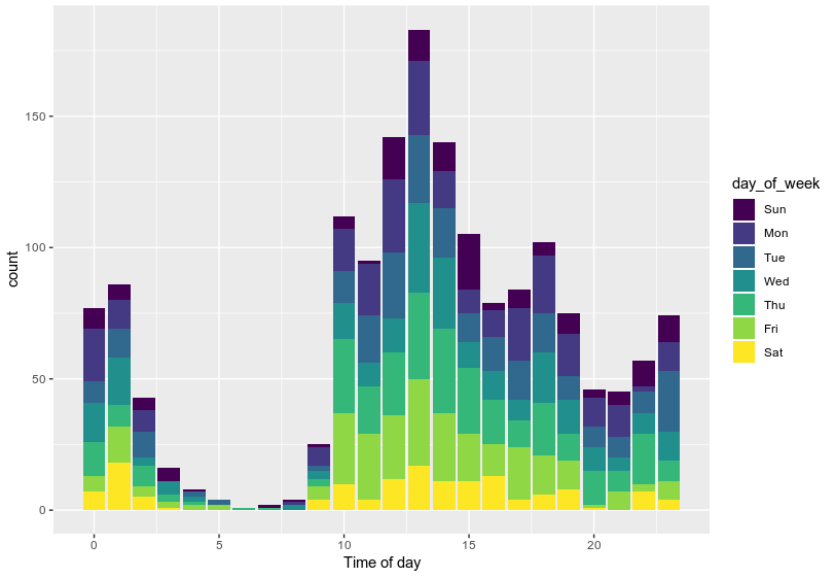

Fig. 6. User activity at a certain time of day with grouping by day of the week.

Determining the activity of the users of the forum has allowed us to identify the necessary timeframes for active deployment of discussion on web forum pages, to model user behavior and analyze it. It also helped to find certain correlations between the time of day, day of the week and the number of generated posts in the content of the historical web forum. Consequently, most users are actively posting new posts on weekdays, especially during the day. Many of them join the discussion on the forum, only when it appeared in a sufficient number of posts (more than half). This allows you to find a good time to collect data, as well as participate in discussions in historical web forums.

4 Conclusions

Modern web forums generate unique sources of historical information that contain large volumes of important, valuable information from eyewitness events. Investigating such a large array of unstructured historical information through web forums is an additional tool of cognition for historians. Web forums provide wide access to information that has not been filtered. Web Mining technology provides a high-quality analysis of the content of web-forums and allows you to get new information about historical events, to conduct operational monitoring of data. Using interpreted object-oriented programming language Python helps to automate the following processes: data collection, identification of the main characteristics of the individual structural elements of the forum, data acqusition, data computing and processing. The tools used in the research provided for the rapid processing of tabular data and helped to identify the most interesting characteristics of historical information that was generated in the content of web forums.

References

1. Trunzer, E., Kirchen, I., Folmer, J., Koltun, G., Vogel-Heuser, B.: A flexible architecture for data mining from heterogeneous data sources in automated production systems. In: 2017 IEEE International Conference on Industrial Technology, ICIT 2017, pp. 1106–1111. Toronto, Canada (2017).
2. McKenna, B., Myers, M. D., Newman, M.: Social media in qualitative research: Challenges and recommendations. Information and Organization 27(2), 87–99 (2017).
3. Zeng, M. L.: Smart data for digital humanities. Journal of data and information science 2(1), 1–12 (2017).
4. Khymytsia, N., Ustiyanovich, T.: Application of Big Data in Historical Science. In: Proceedings 7th International Academic Conference of Young Scientists “Humanities and Social Sciences 2017”, HSS 2017, pp. 368–370. Lviv (2017).
5. Zhezhnych, P., Khymytsia, N., Lisina, S., Morushko, O.: Analysis of computer-based methods for processing historical information. In: Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, pp. 365–368. Lviv (2017).
6. Artem, K., Kunanets, N., Holoshchuk, R., Pasichnik, V., Rzheuskyi, A.: Information Support of the Virtual Research Community Activities Based on Cloud Computing. In: Proceedings of the 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018, pp. 199–202. Lviv, Ukraine (2018).
7. Miller, K.: Communication Theories: Perspectives, processes, and contexts. 2nd edn. McGraw-Hill, New York (2005).
8. Rheingold, H.: The virtual community: Homesteading on the electronic frontier. Addison-Wesley Publishing Company, Reading, MA (2000).
9. Christakis, N. A.: Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives – How Your Friends’ Friends’ Friends Affect Everything You Feel, Think, and Do. In: Christakis, N. A., Fowler J. H. (eds). Back Bay Books (2011).
10. How to: Manage a Sustainable Online Community, https://mashable.com/2010/07/30/sustainable-online-community/.
11. Yalan, Y., Xianjin, Z., Jinchao, Z., Xiaorong H.: Comparing digital libraries with virtual communities from the perspective of e-quality. Library Hi Tech 32(1), 173–189 (2014).

12. Peleshchyshyn, A., Khymytsia, N.: Historiographic aspect of giving information about the situation in the ATO area in virtual communities. In: Proceedings of the 4th International Scientific Conference “Information, Communication, Society”, ICS 2014, pp. 238–240. Lviv (2014).
13. Khymytsia, N.: Socio-focused online research sources Eurorevolution 2013-2014 in Ukraine. The state and the army 784, 214–223 (2014). (in Ukrainian).
14. Fedushko, S.: Development of a software for computer-linguistic verification of socio-demographic profile of web-community member. Webology 11 (2), article 126 (2014).
15. Korobiichuk, I., Fedushko, S., Juś, A., Syerov, Y.: Methods of Determining Information Support of Web Community User Personal Data Verification System. In: Szewczyk R., Zieliński C., Kaliczyńska M. (eds) Automation 2017. Advances in Intelligent Systems and Computing, vol. 550, pp 144–150. Springer (2017). DOI: 10.1007/978-3-319-54042-9_13.
16. Trach, O., Vus, V., Tymovchak-Maksymets O.: Advanced search query for identifying Web-forum threads relevant to given subject area. In: 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET 2016, pp. 849–852. Lviv (2016).
17. Bilushchak T., Peleshchyshyn A., Komova M.: Development of method of search and identification of historical information in the social environment of the Internet. In: XIth International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, pp. 196–199. Lviv (2017).
18. Bilushchak, T., Myna, Zh., Yarka, U., Peleshchyshyn, O.: Integration processes in the archival section of Lviv Polytechnic National University. In: 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, pp. 200–203. Lviv (2017). DOI: 10.1109/STC-CSIT.2017.8098768.
19. Khomytska, I., Teslyuk, V., Holovatyy, A., Morushko O.: Methods, models and means of the system for differentiation of phonostatistical structures of engglish functional styles. Development of methods, models and means of authorship attribution of a text. East European Journal of Advanced Technologies 3(2), 41–46 (2018). DOI: 10.15587/1729-4061.2018.132052.
20. Khymytsia, N., Lisina, S., Morushko, O., Zhezhnych, P.: Peculiarities in generating historical information in virtual communities. In: Proceedings of the 12th Interna-tional Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 1, art. no. 8098799, pp. 336–339. Lviv (2017)
21. Chen, Y. R., Chen, H. H.: Opinion spam detection in web forum: a real case study. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015, pp. 173–183. Florence, Italy (2015).
22. Zhezhnych, P., Markiv, O.: A linguistic method of web-site content comparison with tourism documentation objects. In: Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, pp. 340–343. Lviv (2017). DOI: 10.1109/STC-CSIT.2017.8098800.
23. Shumway, R. H., Stoffer, D. S.: Time series analysis and its applications: with R examples. Springer (2017).
24. Bernaschina, C., Brambilla, M., Mauri, A., Umuhoza, E.: A Big Data Analysis Framework for Model-Based Web User Behavior Analytics. In: Cabot J., De Virgilio, R., Torlone, R. (eds.) Web Engineering, ICWE 2017, vol 10360. Springer, Cham (2017).