Subject Area Study: Keywords in Scholarly Article Abstracts Graph Analysis

Anastasiia Chernysheva[0000-0002-9956-6607], Maksim Khlopotov[0000-0002-9053-027X] and Dmitrii Zubok[0000-0002-2550-7081]

ITMO University, 49 Kronverkskiy pr., lit. A, 197101 St. Petersburg, Russia
avchernysheva@itmo.ru, khlopotov@itmo.ru, zubok@itmo.ru

Abstract. This paper presents an approach to subject area study based on keywords extracted from scholarly article abstracts graph analysis. Initial case study – Digital Humanities, data source – Google Scholar, time interval – 2013–2019. The study is in two parts. First, keywords and key phrases extraction algorithm based on the combination of four existing methods is proposed. The accuracy is up to 77% as we apply strict restrictions to the algorithm thus obtaining better results than other existing solutions provide when are being applied to such short texts as abstracts. Second, keywords graph is created, and its analysis is performed. Applied here graph theory gave an opportunity to detect the most valuable nodes – keywords – along with subareas and closely related areas, showed some trends in Digital Humanities development. Further research proved our approach applicability to other subject areas and data sources.

Keywords: Computational Linguistics, Keywords Extraction, Graph Theory, Subject Area Study, Digital Humanities.

Introduction

Digital Humanities is a new, rapidly developing field, which is gradually becoming a subject of interest for Russian scientists and researchers. So far, this area of knowledge is believed to be represented mostly by natural language processing and data visualization. However, the full range of areas covered by or closely related to Digital Humanities, is not specified [1]. The original idea was to study Digital Humanities as a subject area by extracting and analyzing keywords from Google Scholar scientific data.

Keywords graph represents the approximate subject area structure and makes it possible to work out curriculums, to expand them in accordance with the most relevant scientific trends. As a result, to come up with the research ideas within the field of interest and figure out some directions the subject area is going on thus providing researches with fresh ideas and pointing out out-of-date or already studied enough topics. In addition, the creation of such a graph makes it possible to track the dynamics of the subject area development and, in the future, even predict it.

The paper is structured as follows. Section 2 reviews related works. Section 3 provides an overview of the keywords extraction process and algorithm proposed in this paper. Section 4 emphasizes on keywords graph creation and analysis. Section 5 describes subject area study approach, the highlight of the paper. Finally, Section 6 discusses the achieved results and concludes the work.

1 Related Work

Subject area analysis methods based on graph theory and keywords extraction has been studied previously and there are similar solutions applicable under specific conditions. L. Weston et al. describe an approach to materials science analysis. They apply text mining with named entity recognition (NER) for large-scale information extraction from the published materials science literature. The NER model is trained to extract summary-level information from materials science documents. The result is represented in a structured format, usually graph-structured [2].

Jefferson de J. Costa et al. propose a way to represent undergraduate programs as a directed acyclic graph (DAG), in which each course is represented as a node, and relations between courses are represented as edges. They proposes methods for mining DAGs using statistical analysis and apriori-based concepts, to identify retention patterns in undergraduate programs [3].

O. Faust made a review on promoting the use of computing machinery by the *Computers in Biology and Medicine* journal in the fields of bioscience and medicine. Analysis of the author supplied keywords was carried out. Keywords clustering showed the statistical connection between them and helped to identify the most popular topics and trends. The results were visualized with graphs [4].

H. Sekiguchi et al. analyzed the guidelines of the American Heart Association Basic Life Support using data mining methods to identify and characterize the changes in keywords and key points. They also built and analyzed a co-occurrence network to classify the words into major topics on one step of the research [5].

Y. Solomonova and M. Khlopotov present an approach to Russian text vectorization based on SRSTI classifier. They use keywords extraction to define SRSTI categories as lists of keywords. The keywords selection process is described, and vector calculation and comparison algorithm are applied to marked-up SRSTI texts [6].

Sharma et al. propose a topic network analysis approach using topic modeling and network analysis. They carried out an experiment on the field of Machine Learning and detected main topics and trends in the area along with interrelationships [7].

Subject area keywords graph creation appears to be a one-size-fits-all solution for the formal subject area analysis, which gives an opportunity to understand its structure. Existing methods and approaches for analyzing subject areas are often field-specific and are not applicable outside of one or more subject areas. It is worth mentioning that in some of them the use of graph theory for the subject area analysis is proposed. However, it differs radically from the approach proposed in this paper. A more thorough review of the methods for subject area analysis showed, on the one hand, their general non-universality, and on the other, the need to engage an expert.

2 Keywords Extraction

To analyze subject area in terms of its structure its elements and subareas should be determined. It can be done by analyzing the scientific literature (in this study, it was decided to work with articles) and identifying keywords.

There are two ways to extract keywords from text:

1. With the engagement of specialists in the studied area.
2. Using algorithms for automatic keywords extraction.

In this research, both methods were discussed, and a conclusion was drawn on the inappropriateness of applying the expert approach, which determined the main stages of the study.

The implementation of the study was carried out in the Python programming language using:

1. scholarly, a Python-based module that allows to retrieve author and publication information from Google Scholar [8].
2. PKE, an open source Python-based key phrase extraction toolkit. It provides an end-to-end keyphrase extraction pipeline in which each component can be easily modified or extended to develop new models [9].
3. Yandex.Translate, a Python module for Yandex.Translate API [10].
4. NLTK, Natural Language Toolkit, a platform for building Python programs to work with human language data [11].
5. Scikit-learn, efficient Python-based tool for data analysis and machine learning [12].

2.1 Source Selection

Google Scholar, a part of the Google search engine, was chosen as a source of scientific materials. Unlike other sources such as Scopus, Web of Science and IEEE, it provides free access to the highest possible number of scientific papers from all over the world in different languages from peer-reviewed journals.

In addition to scientific works, with the help of Google Scholar information about researchers, their scientific interests, authors' citations, and publications can be retrieved. This study is focused on information on scientific articles and their authors, as it is possible to form an idea of the subject area itself according to the leading researchers' lists of interests. The only significant Google Scholar's drawback is that it does not provide access to the author's keywords to the article, which makes it necessary to extract keywords from the abstract body.

2.2 Data Collection

In the view of data collection from Google Scholar peculiarities and the needs of the study, it was decided to collect the following data about the authors:

- Name.
- Affiliation.
- Citations.

- Scientific interests.

Data was collected on the profiles of 1,106 authors – all who put Digital Humanities in the list of their research interests.

With respect to the publications, it was decided to collect the following data:

- Author(s).
- Title.
- Publication year.
- Journal where the article was published.
- Abstract.

Data was collected on 13,847 publications.

2.3 Data Pre-processing

Firstly, authors' list of interests pre-processing was performed. It contained 4,535 terms. To analyze keywords successfully, it was necessary to solve the following problems:

Authors write down their interests in different languages, therefore, they need to be translated into one language, for convenience – English, as 11,907 articles are in English (79% of the total number of articles). 42 different languages were detected within the collected data.

One term can be written down in different ways (abbreviations or in full, include typos, etc.), for example, "data visualization" was detected written down in nine different ways:

1. Data visualization.
2. Visualisation.
3. Visualization.
4. Data visualisation.
5. Information visualization.
6. Information visualisation.
7. Metadata visualization.
8. Metadata visualisation.
9. Datavis.

Thus, the list of keywords was pre-processed in seven steps. The examples illustrate keywords from authors' interests pre-processing.

1. Automatic keywords translation into English. Figure 1 shows the results of several words translations.

2. Automatic translation errors manual correction.

Example: el siglo de oro (Spanish) → century of gold → golden age.

There were 135 non-English unique terms, 27 were translated incorrectly.

3. Terms writing standardization.

Example: vr, virtual world, virtual reality → virtual reality.

4. Separation of different terms united by "and / &".

Example: augmented and virtual reality → augmented reality, virtual reality.

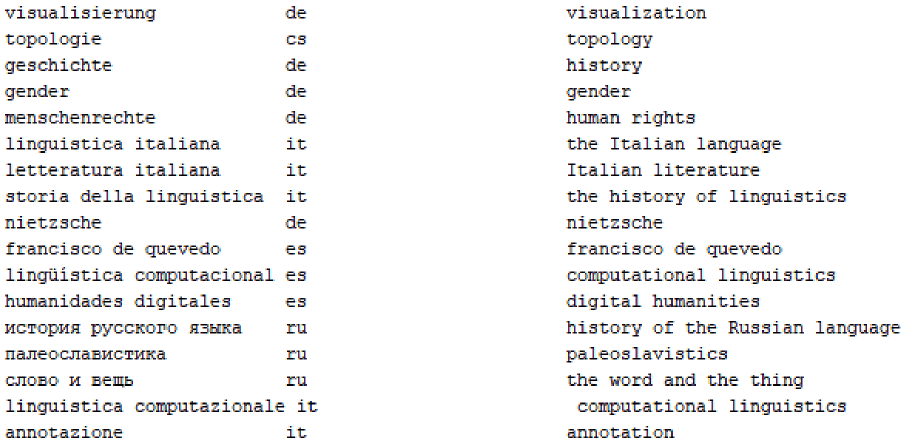
5. "The" and "a" removal.

6. Removal of “interests” that do not make sense for this study, for example “coping with life stress”, “I research”.

7. Encoding bugs removal.

As a result of processing, a list of 4,445 words was obtained.

Similar keywords pre-processing was performed on the list of words extracted from abstracts using regular expressions.



visualisierung	de	visualization
topologie	cs	topology
geschichte	de	history
gender	de	gender
menschenrechte	de	human rights
linguistica italiana	it	the Italian language
letteratura italiana	it	Italian literature
storia della linguistica	it	the history of linguistics
nietzsche	de	nietzsche
francisco de quevedo	es	francisco de quevedo
lingüística computacional	es	computational linguistics
humanidades digitales	es	digital humanities
история русского языка	ru	history of the Russian language
палеославистика	ru	paleoslavistics
слово и вещь	ru	the word and the thing
linguistica computazionale	it	computational linguistics
annotazione	it	annotation

Fig. 1. Automatic translation example

2.4 Keywords Extraction Algorithm

First, existing algorithms for automatic keywords extraction analysis was carried out. More than 19 algorithms were discussed, nine tested. During testing, algorithm requirements were defined:

- Mainly nouns or phrases where the main word is a noun, and the definitive - adjectives, participles or less often adverbs, should represent keywords (TF-IDF [13]: “labeled”, “using”).
- A pronoun cannot be a part of a key phrase (PositionRank [14]: “our method”).
- Single adjective cannot be considered a keyword, adjectives can only be a part of a key phrase where the main word is noun (KP-Miner [15]: “efficient”, “beautiful”).
- Long phrases and whole sentences cannot be key phrases.
- A key phrase should not be incomplete (Rake: “efficiently map text”, “online procedure used”, YAKE [16]: “classification of multi”).
- List of key terms should not be represented only by single keywords or only by composite key phrases.

Finally, we have selected four algorithms, all implemented by NLTK and PKE libraries:

- TF-IDF.
- TextRank [17].
- PositionRank.

– MultipartiteRank [18].

These four algorithms were used as a basis for our algorithm, which demonstrated higher keywords extraction accuracy in terms of our task and conditions.

Algorithm development. To describe an algorithm which meets the requirements and is based on the tested and optimized combination of four mentioned above, set theory is used.

$$A = PositionRank \cap MultipartiteRank \tag{1}$$

Such an intersection gives a stable set of key phrases. As PositionRank and MultipartiteRank tend to extract key phrases, not keywords, single keywords are potentially lost, so other algorithms results should be considered. Nevertheless, *A* may include set of single keywords *S*, so

$$A = S \cap F, \tag{2}$$

where *F* is a set of key phrases.

$$B = TF\text{-}IDF \cap TextRank \tag{3}$$

It was experimentally established that TF-IDF and TextRank algorithms (3), highlighting mostly single keywords, tend to select incomplete phrases and verb constructions as key phrases, most of which are lost after intersection

$$B = S1 \cap F1, \tag{4}$$

where *F1* and *S1* are sets of keywords and key phrases of set *B*, respectively.

$$C = (B \setminus S) \cup (B \setminus F1) \tag{5}$$

In (5) filtering out the set obtained in (3) is performed.

$$C1 = B \setminus S \tag{5.1}$$

In (5.1) single keywords occasionally included into *A* are excluded from *B*.

$$C2 = B \setminus F1 \tag{5.2}$$

In (5.2) all key phrases are excluded from *B*. In (6) sets of keywords and key phrases are being united.

$$D = A \cup C \tag{6}$$

The algorithm is configured so that verbs and adjectives cannot be single keywords, and PositionRank and MultipartiteRank key phrases does not include verbs due to part of speech constraints. Despite these facts, due to English words ambiguity verb still may be considered a keyword.

The problem of extracting key phrases containing pronouns was solved by expanding the list of stop words.

Testing results evaluation. Table 1 shows the results of nine selected algorithms testing in comparison with the developed one, where KW – keywords, KP – key phrases, IP – incomplete phrases, SA – single adjectives, V – verbs, P – pronouns. Testing was carried out on a 100 randomly selected abstracts previously marked up by experts. Here, numbers represent the percentage of words for each category compared to the total number of keywords and key phrases extracted from each abstract; the average values by the 100 abstracts are given.

The table shows that it was not possible to eliminate the inclusion of verbs and adjectives, as well as incomplete phrases in the list of article abstract keywords, although their number was significantly reduced. Precision, recall and F-measure for a set of 100 random abstracts equal 76.3%, 52.6% and 62.27%, respectively. The precision of keywords extraction is of higher importance than its recall, which was acceptable within the study. In addition, too many keywords might negatively affect the resulting graph.

Diagram in the Figure 2 illustrates the accuracy of nine algorithms that showed the best results during testing, as well as the algorithm we developed. Here, we note that 76.3% is a low accuracy for a keyword extraction algorithm under normal conditions, and most of the tested algorithms demonstrate better results when larger texts are being processed.

However, given the small size of abstracts (3–6 sentences), the low frequency of keywords within the abstract body, and the non-semantic approach of automatic algorithms, the obtained accuracy is considered high.

Table 1. Keywords extraction performed by our and 9 other algorithms

Algorithm	**KW**	**KP**	**IP**	**SA**	**V**	**P**
TF-IDF, %	80	20	12.6	16	34	0
KP-Miner, %	69	31	11	15.3	31.2	5.6
Rake, %	19	81	46.4	8	28.6	0.4
YAKE, %	67	33	13.7	40.8	6	0
TopicRank, %	67	33	4.7	37	7.8	0.9
TopicalPageRank, %	14.3	85.7	15.7	12.1	13.2	13.4
PositionRank, %	27	73	0	0.9	0.9	14.8
MultipartiteRank, %	47	53	0	6	0	10.4
TextRank, %	53	47	7	0	3.4	0
Our method, %	28	72	0.7	3.7	2.9	0

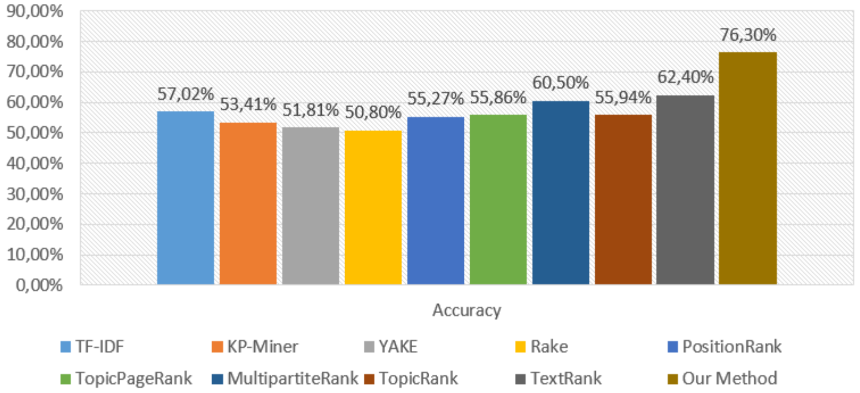


Fig. 2. Algorithms' accuracy

The developed algorithm has been tested on scientific publications abstracts, news articles abstracts, and full-text scientific and news articles. Keywords extraction from essays, fiction and conversational texts was not carried out.

The extraction of keywords from news articles abstracts was performed with no less accuracy than when working with scientific data, while keywords extraction from full-text papers turned out to be almost inapplicable. The results obtained indicate that the developed algorithm is field-focused. This can be explained by the fact that in our task it was decided that extracted keywords' quality prevails over their number and the length of processed texts usually does not exceed 1,500 characters – the approximate size of an abstract. The developed algorithm has strict constraints and is not optimal for other tasks.

3 Keywords Graph Creation and Analysis

91,447 words were extracted from the abstracts; 50,962 of them are unique, which is approximately 56% of the total number of extracted keywords. Top-20 keywords are presented in Table 2.

Keywords graphs were created and analyzed both for the researchers' scientific interests and for keywords extracted from abstracts. Each keyword or key phrase is a node. If two keywords are extracted from one abstract, we consider there is a connection of unknown type between the keywords and create an edge. The same goes for keywords from the researchers' scientific interests.

In the study, a graph of keywords obtained from abstracts' keywords is of greater interest. Its size (in terms of the number of nodes and edges) is comparable to the size of a social graph, so it made sense to use similar approaches to its analysis.

The intersection of the central nodes sets obtained by calculating betweenness centrality, eigenvector centrality and degree for each node gives a stable set of central and most significant nodes of the keywords graph that determines the subareas of the subject area.

Table 2. Top-20 Digital Humanities keywords based on its occurrence

№	Keyword	Occ.	№	Keyword	Occ.
1	natural language processing	**57**	11	archaeology	**24**
2	data visualization	**49**	12	information retrieval	**21**
3	computational linguistics	**41**	13	artificial intelligence	**21**
4	librarianship	**34**	14	text mining	**20**
5	information technologies	**34**	15	book history	**20**
6	history	**29**	16	cultural heritage	**20**
7	digital library	**29**	17	geographic information system	**20**
8	media studies	**28**	18	library	**19**
9	literature	**27**	19	human-computer interaction	**19**
10	machine learning	**27**	20	scholarly communication	**19**

After that, community detection is performed. This type of clustering refers to the procedure of identifying groups of interacting nodes in a graph depending upon their structural properties [19]. Community detection in graphs allows to combine keywords related to one subarea. Further, centralities calculation performed for each community shows which nodes determine the main topic of the cluster and which neighboring subareas are the most closely related to the one of interest.

Due to the substantial number of nodes (50,961) and edges (217,438), and the importance of each node rather than a group of nodes, clustering was carried out in the first place. Resulting graph was of 89 communities, considered separately later.

Figure 3 shows the cluster of Natural Language Processing (in English), and Figure 4 shows the graph of all Digital Humanities subareas (in Russian), obtained by analyzing 89 selected communities.

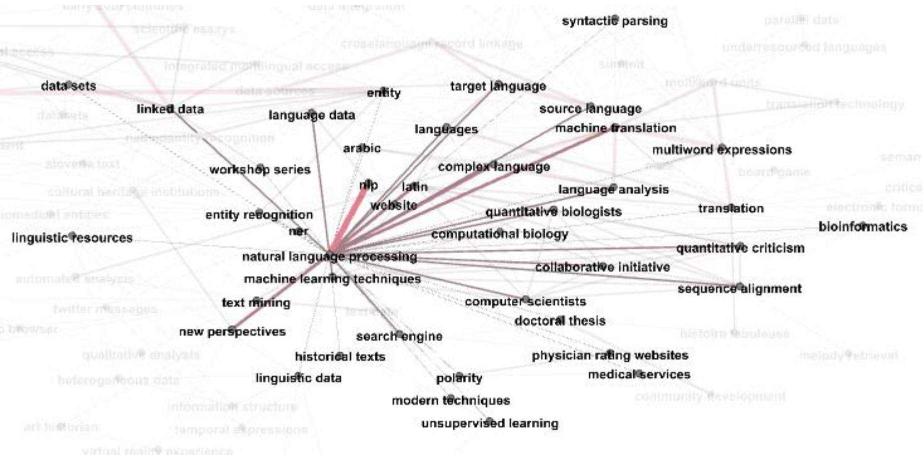


Fig. 3. Natural language processing cluster (neighboring keywords)

The subareas graph has 75 nodes and 118 edges. Due to the uneven clustering of the graph, it would be incorrect to draw conclusions about the nodes' centralities, however, the graph shows connections between the Digital Humanities subareas, and clearly identifies nodes that have the greatest number of connections with others, namely:

- History.
- Text Analysis.
- Machine Learning.
- Natural Language Processing.
- Data Analysis.
- Information Technologies.

According to the graphs created separately for each of the past three years, changes in the research areas in the field of Digital Humanities can be traced. At first, only the humanities were in the foreground, such as Linguistics, History, Archeology and Data Visualization. By 2019, Information Technologies, Machine Learning, Databases are of greater importance, and here appears a large cluster associated with Medical research. This indicates a rapid expansion of the subject area, as well as it shows that scientists of various fields are rapidly becoming interested in Digital Humanities.

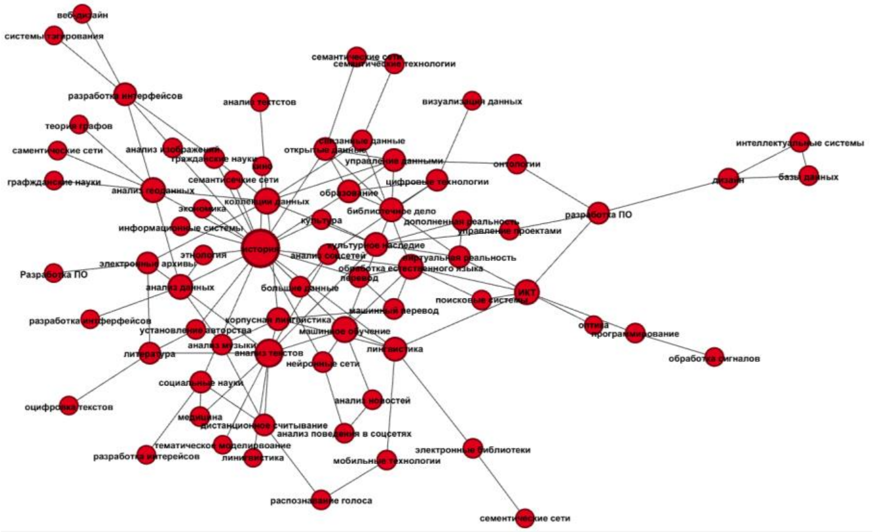


Fig. 4. Subareas graph

4 Subject Area Study Approach

The proposed approach has three main steps:

- Scientific materials collection.
- Keywords extraction from article abstracts.
- Keywords graph creation and analysis.

In Figure 5 the approach is described in detail.

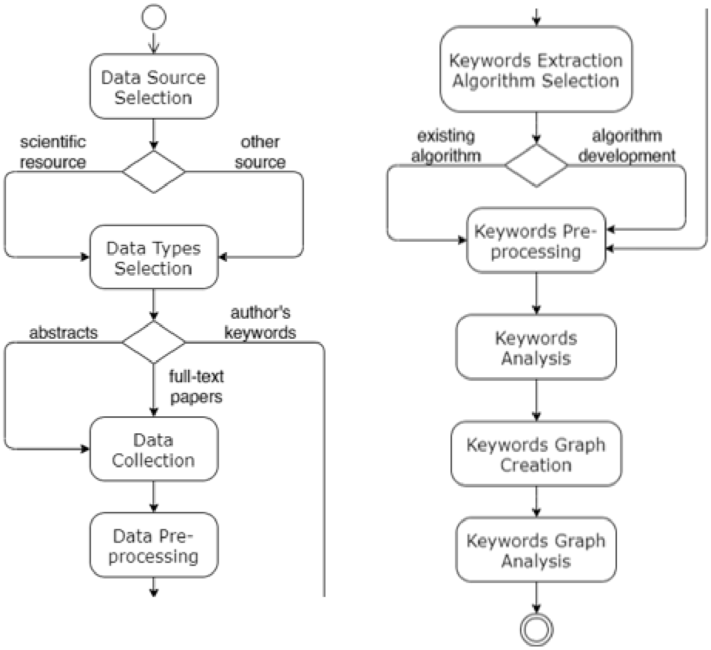


Fig. 5. Subject Area Study Approach

Conclusion and Future Work

In this paper, an approach to subject area analysis based on scholarly article abstracts keywords graph creation is proposed; its steps are described in detail for Digital Humanities area. Data was collected from Google Scholar, though additional research proved that the approach is applicable to other data sources and other subject areas as it was tested on arXiv.org data, and Elsevier Scopus and ScienceDirect data and applied to other subject areas, namely, Multimedia, Databases and Machine Learning.

Keywords extraction algorithm was described. Its accuracy is up to 77% which is quite a high result in terms of our study and conditions. Keywords graph analysis was represented by discovering subareas, graph communities, detecting most important ones and applying an idea of analyzing trends in subject area development.

Our future work will be focused on determining types of connections between keywords (nodes) as it widens the scope of research along with russification of the proposed keywords extraction algorithm. The latter seems to be more difficult task as there are not so many approaches to work with texts written in Russian. As we go forward, we plan to use more up-to-date natural language processing and machine learning methods in order to obtain higher accuracy of the proposed keywords extraction algorithm.

References

1. Digital Humanities, https://ling.hse.ru/Projects_DigHum, last accessed 2019/06/15.
2. Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K.A., Ceder, G., Jain, A.: Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. Journal of Chemical Information and Modeling 59(9), 3692–3702 (2019).
3. Costa, J.D.J., Bernardini, F., Artigas, D., Viterbo, J.: Mining direct acyclic graphs to find frequent substructures — An Experimental Analysis on Educational Data. Information Sciences 482, 266–278 (2019).
4. Faust, O.: Documenting and Predicting Topic Changes in Computers in Biology and Medicine: A bibliometric Keyword Analysis from 1990 to 2017. Informatics in Medicine Unlocked 11, 15–27 (2018).
5. Sekiguchi, H., Fukuda, T., Tamaki, Y., Hanashiro, K., Satoh, K., Ueno, E., Kukita, I.: Computerized data mining analysis of keywords as indicators of the concepts in AHA-BLS guideline updates. The American Journal of Emergency Medicine, 38 (7), 1436-1440 (2019).
6. Solomonova, Y., Khlopotov, M.: Russian Text Vectorization: An Approach Based on SRSTI Classifier. Communications in Computer and Information Science 1038, 754–764 (2019).
7. Sharma, D., Kumar, B., Chand, S.: Trend Analysis of Machine Learning Research Using Topic Network Analysis. Communications in Computer and Information Science, 34–47 (2018).
8. Scolarly, https://github.com/OrganicIrradiation/scholarly, last accessed 2020/02/02.
9. Boudin, F.: PKE: An Open Source Python-Based Keyphrase Extraction Toolkit. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, 69–73 (2016).
10. Python module for Yandex.Translate API, https://github.com/dveselov/python-yandex-translate, last accessed 2019/12/18.
11. Natural Language Toolkit, https://www.nltk.org, last accessed 2019/11/13.
12. scikit-learn: machine learning in Python – scikit-learn 0.19.1 documentation, Scikit-learn.org, http://scikit-learn.org/stable/#, last accessed 2019/11/13.
13. Robertson, S.: Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation 5, 503–520 (2004).
14. Florescu, C., Caragea, C.: PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1105–1115 (2017).
15. El-Beltagy, S., Rafea, A.: KP-Miner: Participation in SemEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, ACL, pp. 190–193 (2010).
16. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., Jatowt, A.: YAKE! Collection-Independent Automatic Keyword Extractor. In: Advances in Information Retrieval (2018).
17. Understand TextRank for Keyword Extraction by Python, https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0, last accessed 2019/02/20.
18. Boudin, F.: Unsupervised Keyphrase Extraction with Multipartite Graphs (2018).
19. Yang, J., McAuley, J., Leskovec, J.: Community Detection in Networks with Node Attributes. In: Proceedings of IEEE International Conference on Data Mining, ICDM, pp. 1151–1156 (2013).