

Keyword-Based Comparison of Scientific Databases

Anna Cena¹, Iryna Balagura^{2,3}

¹ *Warsaw University of Technology, Koszykowa Street, 75, 00-662 Warsaw, Poland*

² *Institute for Information Recording of National Academy of Sciences of Ukraine, 2, Mykolya Shpaka Street, Kyiv, 03113, Ukraine*

³ *University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom*

Abstract

Keyword analysis is a widely used technique which includes statistical and co-word analysis that allows study of the development of research domains, identification of hot topics and new trends, prediction of knowledge evolution, and investigation of the interdisciplinary nature of science concepts. The main aim of the research presented in this paper is to better understand the mechanisms governing the popularity of specific topics and how they change over time. In the paper, the open-access databases *ArXiv* and *DBLP*, and the *Stack Exchange Q&A* websites were compared using keyword analysis in the computer science area. The most popular topics in computer science were detected using time-series and co-word network. The behaviour of keywords and the mechanisms governing the popularity of specific topics were investigated.

Keywords

scientific databases, *DBLP*, *ArXiv*, *StackExchange*, bibliometrics, keywords

1. Introduction

Co-word network analysis is frequently used in bibliometrics because it provides a clear graphic visualization, helps to describe the structure of the subject area, and highlights the main key elements and groups of keywords to identify subtopics and interdisciplinarity in science. Moreover, this perspective opens up many opportunities for additional analysis, e.g. with tools developed within a theory of complex networks. Please note that a combination of co-author and co-word networks allows the construction of a heterogeneous information network. In [1] a Meta Path Computed Prediction (MPCP) algorithm for link prediction among scientists and publications was presented. Authors of [2] used co-word network modularity analysis to identify primary research interests. The development of scientific areas also includes the comparison analysis of domains: Khajavi et al. in [3] proposed to measure a fuzzy distance between two domains using the three indicators of frequency, development, and investment appeal.


Keyword analysis helps to observe the rise and fall of scientific concepts; for example, in [4] it was shown that fields consistently follow a rise and fall pattern captured by two parameters of right-tailed Gumbel temporal distribution. Keyness analysis is used to identify significant keywords in different time periods by comparing the difference between the observed frequencies and the expected frequencies [5].

A keyword analysis is used for identifying hot topics using the most frequent terms in a particular domain. Hot topics refer to “issues and topics that are discussed by a relatively large number of scholars within a certain period” [6]. For instance, Park et al. in [7] proposed a keyword-scoring metric to measure the degree of the emergence of a word compared to the terms in a particular domain. The same principles are included for new and emergent trend detection. By analyzing time series keywords frequency on time intervals, it is possible to detect new or emerging topics [8]. In some cases, additional evidence, such as investment, could support co-word network analysis to provide an

ITS-2023: Information Technologies and Security, November 30, 2023, Kyiv, Ukraine

✉ anna.cena@pw.edu.pl (A. Cena); balaguraira@gmail.com (I. Balagura)

ORCID [0000-0001-8697-5383](https://orcid.org/0000-0001-8697-5383) (A. Cena); [0000-0001-9627-2091](https://orcid.org/0000-0001-9627-2091) (I. Balagura)

 © 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

evaluation of research topics. Authors of [9] suggested the use of a keyword co-occurrence network together with investment, measured by the number of sponsors associated with each keyword.

Co-word network analysis is the most frequently used instrument in bibliometric analysis because it provides a clear graphic visualization, helps to describe the structure of the subject area highlights the main key elements and groups of keywords to identify subtopics and interdisciplinarity in science, opens up many opportunities for additional analysis according to the developed theory of complex networks, and provides opportunities for new applications of already existing methods. For instance, Lande et al in [10] conducted keyword analysis using a co-occurrence network of categories of subject domains and their density, to identify related topics in scientific research, detect trends in research, search for interdisciplinary terminology and correct usage of terms, and describe science structure.

The main aim of the research presented in this paper is to better understand the mechanisms governing the popularity of specific topics and how they change over time. The investigation carried out in this paper includes a comparison of selected open-source bibliometric databases, i.e., DBLP database and *ArXiv* e-print database and *Stack Exchange Q&A* websites, in terms of keyword behaviour. We are interested in similarities and differences between these data sources as well as the information concerning relationships between topics and their evolution, interdisciplinarity and possible predictions for the future that can be formulated based on them. In the example of the computer science category/research field of study, various modelling techniques are considered, from quantitative analysis to keyword co-occurrence network modelling.

2. Scientometric databases

In this section databases, namely *DBLP*, *ArXiv*, and *Stack Exchange*, will be described and compared. We focused on open sources that allow bibliometric information to be gathered and analyzed freely.

Many studies have been conducted on the qualitative and quantitative comparison of several scientific databases [11,12]. But the lifecycle of technology in applied sciences includes also development and usage in real life. We assume that the analysis and comparison of academic and non-academic databases allow us to find the main trends in applied sciences.

The *DBLP* database is a widely known computer science abstract database, which indexes approximately 7M publications, over 3M authors, over 6K conferences, and approximately 2K journals. Initially, *DBLP* started as a database systems and logic programming (*DBLP*) research group at the University of Trier in Germany, and since 2018 it has been operated and maintained by Schloss Dagstuhl [13]. *DBLP* is an open-access source of data, that provides data and a clear description of the form of storing data and updating processes.

The open-access and Rosenfeld et al. [14] found an ‘over-indexing’ of Computer Science publications in *DBLP*, but the number of such records was not significant and the difficulty of establishing boundaries for Computer Science in interdisciplinary research should be taken into account. The *DBLP* creators define their database as not complete but the processes of adding and improving it are ongoing [13]. The *DBLP* is recognized as a reliable database in Computer Science. However, today there is no abstract resource that would provide a complete reflection of at least one scientific field.

ArXiv is a free distribution service and an open-access archive with 2,166,249 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Even though that *ArXiv*’s articles are not peer-reviewed a lot of scientometric research for scoping review in different areas relies on this database in the same range as well-known Scientific databases. In the [15] authors estimated the interdisciplinarity of concepts in data science, in [10] authors studied security development patterns in computer science, and in [16] the scientometric analysis of papers in physics, astronomy and particle physics was presented. The greatest advantage of *ArXiv* is the fast publishing

process, allowing scientists to increase the speed of research exchange, which was critically apparent during the COVID-19 pandemic. Scientometric analysis of papers from *ArXiv* and a gathered dataset from several resources regarding COVID-19 revealed the most active researchers, institutions and most common topic of research in this area in papers [17, 18].

StackExchange is a popular example of a Q&A website. Q&A websites are technical discussion forums on social media that serve as a platform for users to interact mainly via questions and answers and have become a necessary part of professional practice. Thus, the content of such websites mostly consists of current practice topics in different areas. Stack Exchange is a large community run by professionals and enthusiasts and comprises 173 Q&A communities, including *Stack Overflow*. Over 100 million people visit every month to ask questions, learn, and share technical knowledge in different areas [19]. There are a lot of research issues that were solved using the *StackExchange* data in recent years. Authors of papers [20, 21] proposed solutions for improving education through the usage of the *StackExchange*. In the paper [22], online leadership through the linguistic perspective. Similar to scientometric research mining of Q&A websites can discover main communities and change in time of topics of discussion [23]. Q&A websites as sources gave no fewer fruits of research for discovering the area’s development than scientific databases. However, the question of combining the mining of scientific and Q&A resources to obtain the full picture of area development is still open.

All three databases play an important role in the communication process in science and areas of expertise and spreading knowledge. The open policy of the resources makes them more available for uploading information, usage, scientometric and data analysis [24].

3. Empirical analysis

Computer science topics were analyzed and key-words networks were compared based on different resources. Note that in the case of *Stack Exchange* forums, we decided to include both Computer Science and Math forums. Table 1 presents the summarization of data that were included in each of the databases.

For analyzing the DBLP database the v13 dataset was chosen which was released in May 2021 and consists of 5354309 papers and related keywords [25]. Instead of keywords in the databases, *ArXiv* and *StackExchange* categories of records were selected, thus the biggest number of keywords were obtained using dataset DBLP-V13.

Table 1

Characteristics of data sources considered

Database	Timeline	No. of keywords
ArXiv	2007 - 2022	176 ¹
Math ²	2010 - 2022	1930
Computer Science ³	2008 - 2022	652
dblp-v13	1930 - 2022 ⁴	7,299,784

Please note that the datasets from all sources databases were easily accessible (i.e. there are archives available via official channels). Each of them, however, required extensive and time-consuming pre-processing and cleaning. In general, at the first step of the scientometric analysis, data were gathered and filtered. Tags / keywords were extracted (if needed) and standardized. The empirical cumulative distribution function estimated for keyword frequencies in each database is presented in fig. 1., which presents the accumulation of probabilities and supports the data from table 1.

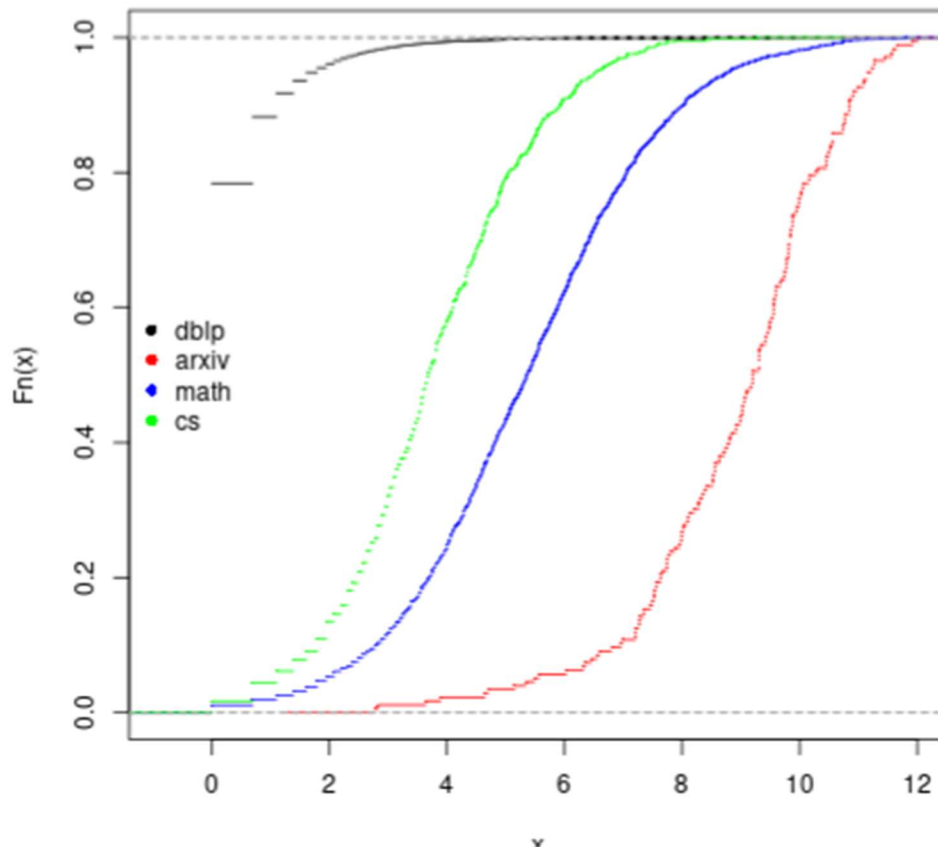


Figure 1: Empirical cumulative distribution function for keywords occurrence.

Next an analysis of the time series of the posts and keywords was performed and co-word networks were built fig.2. For both papers (*DBLP-v13* and *ArXiv* databases) and posts (*Stack Exchange* forums) keyword / tags were mainly used. However, some additional information was also included (e.g. publication date, authors). The usage of keyword fuzzy is presented on fig. 3 as an example. A decline in the number of times this specific keyword was used may be observed. This may indicate for example that the keyword fuzzy was replaced by more specialized and detailed expressions representing this domain.

The keyword networks were built using the frequency of keywords. The presence of keywords/tags in one publication (paper or post) represents the link between nodes in the network. For example, for the *Stack Exchange* forum, the co-occurrence of tags is presented with the network which consists of 664 nodes and 21735 edges, and the network diameter is equal to 4. The nodes with the highest centrality have the most number of connections or are involved in the most number of topics (fig.2). The main keywords for all the data are presented in Table 2.

interdisciplinarity, and a strong connection between computer science and physics.

Table 2

Categories in Arxiv

Categories	Area	Description
hep-ph	High Energy Physics – Phenomenology	Theoretical particle physics and its interrelation with experiment. Prediction of particle physics observables: models, effective field theories, calculation techniques. Particle physics: analysis of theory through experimental results
hep-th	High Energy Physics – Theory	Formal aspects of quantum field theory. String theory, supersymmetry and supergravity
quant-ph	Quantum Physics	No description in Arxiv
cs.LG	Machine Learning	Papers on all aspects of machine learning including also robustness, explanation, fairness, and methodology
astro-ph	Astrophysics	Includes astro-ph.CO(Cosmology and Nongalactic Astrophysics), astro-ph.EP (Earth and Planetary Astrophysics), astro-ph.GA (Astrophysics of Galaxies), astro-ph.HE (High Energy Astrophysical Phenomena), and others

Table 3

Top keywords in computer science

dblp	math	cs	arxiv
data mining	real-analysis	algorithms	hep-ph
feature extraction	calculus	complexity-theory	hep-th
internet	linear-algebra	graphs	quant-ph
computer science	probability	formal-languages	cs.LG
optimization	abstract-algebra	time-complexity	astro-ph

Moreover, the concept with the highest centrality was used as a keyword for searching and gathering the *ArXiv* data. For gathering data from *ArXiv* and *SteckExchange* databases, the *Science Metric library* was used, which was developed by the scientists of the Institute for Information Recording of the National Academy of Sciences of Ukraine [27]. The Science metric library automatically processes the data from several databases such as *Arxiv*, *SteckExchange*, Ukrainian and Chinese abstract databases. The system allows searching by keywords and gives the reports with the time series by the year, most popular keywords in titles and abstracts, and authors forming co-word and co-author networks which could be further analyzed with visualizing software [28].

Networks of concepts related to the selected tags were built using *ArXiv* and *Stack Exchange* and the main characteristics of the networks were calculated. Selected concepts were compared by betweenness centrality measures. Betweenness centrality allows the detection of nodes - connectors of parts of the network, so it is possible to detect the most interdisciplinary keywords. Selected keywords reflect the difference in purpose and usage of the databases [29].

For instance, a search was done for the keyword ‘optimization’ in *Stack Exchange* and *Arxiv*. Fig. 4 shows the co-word network for the dataset obtained by searching words using the keyword ‘optimization’ in the *Science Metric library* in the *Stack Exchange* dataset. The dataset of records from *Stack Exchange* in Computer Science was limited to 2794 records where the word optimization was mentioned. The degree centrality of nodes, which is the number of connections with other keywords, was used to reflect the size of nodes and define the most common keywords. For this network, the keywords with higher degree of centrality were defined such as optimization, constrained-optimization, linear algebra, algorithms, python, convex-optimization, non-linear programming, Matlab, matrix, iterative method, finite-element, quadratic-programming and others. Such keywords reflect the application of optimization in real cases which were discussed in the forum. The clusters

of the network were defined using modularity, which measures the strength of divisions. Three clusters were detected, the largest cluster containing most of the keywords with a higher degree of centrality.

By the same request, *Arxiv* data were analysed and 174167 records were found. The network obtained for the keyword ‘optimization’ with *Arxiv* is shown in fig. 5. This network was not divided into clusters because it consists of one large cluster with the keywords: optimization and control, machine learning, systems and control, numerical analysis, probability, artificial intelligence, learning, neural and evolution, data structures and information theory. This time, the network more precisely describes the connection between fundamental concepts.

The comparison of keywords using betweenness centrality is shown in Table 4.

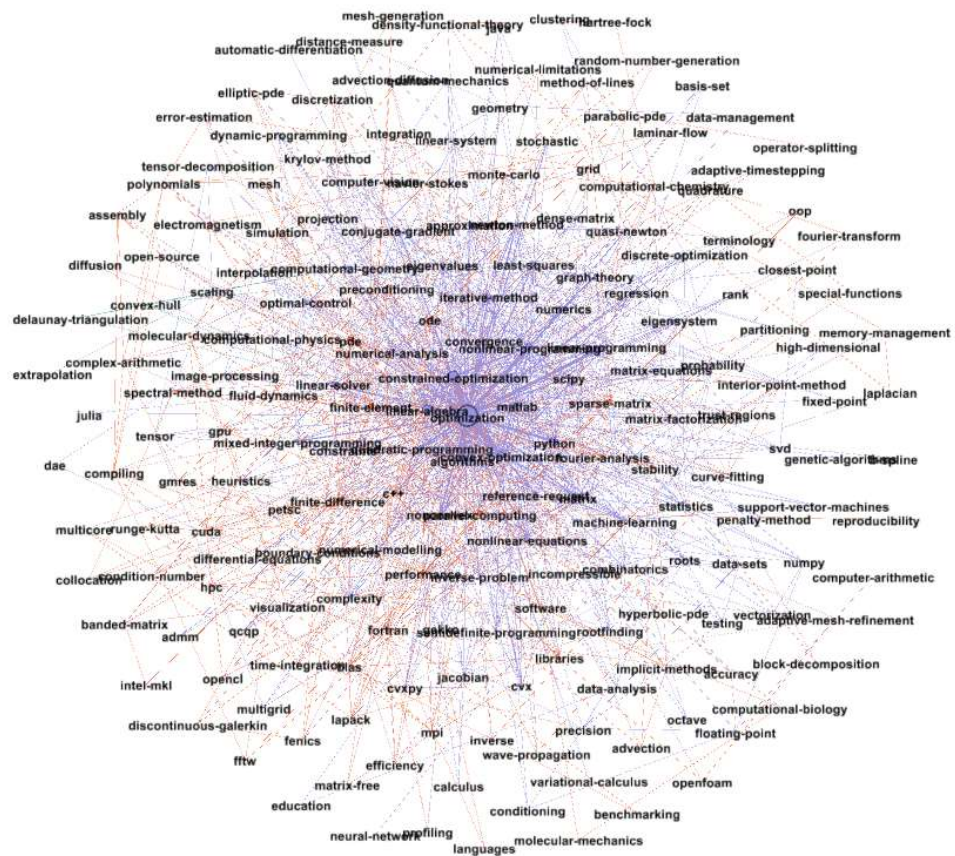


Figure 4: Co-word network for the keyword *Optimization* using data from *SteckExchange*

Table 4

Top keywords in computer science.

ArXiv	Degree	SteckExchange	Degree
Optimization_and_Control	2481.27	optimization	9316.49
Machine_Learning	283.37	constrained-optimization	1366.46
Systems_and_Control	163.41	linear-algebra	1187.42
Numerical_Analysis	76.31	algorithms	819.91
Probability	58.68	python	614.64
Data_Structures_and	53.63	matrix	600.00
Dynamical_Systems	47.33	parallel-computing	462.74
Artificial_Intelligence	35.03	convex-optimization	443.97
Computational_Engineering,_Finance,	25.70	sparse-matrix	442.30
Learning	20.5	pde	431.28

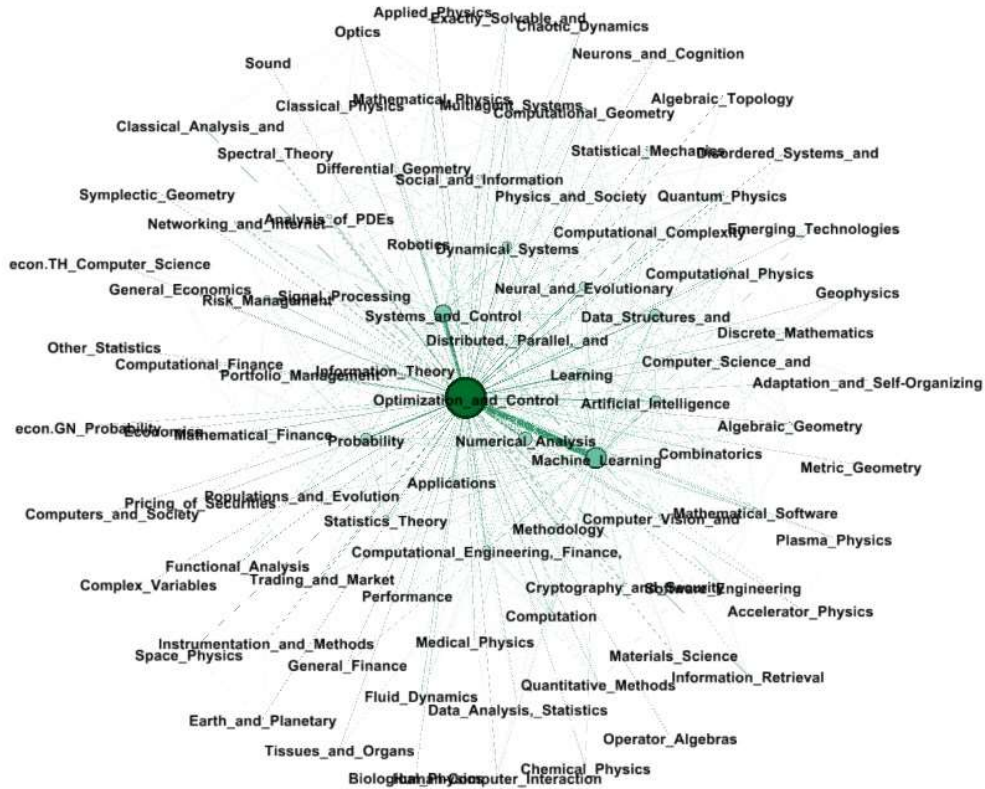


Figure 5: Co-word network for the keyword Optimization using data from *ArXiv*

Using keyword and co-word network analysis, the *Arxiv* database and *SteckExchange* were compared by categories and by the chosen keywords. This allowed us to analyze the datasets on the different levels.

4. Conclusions

By analysing the *DBLP* database, the *ArXiv* e-print database and the *Stack Exchange Q&A* websites, the behaviour of keywords and the mechanisms governing the popularity of specific topics were investigated. The most popular research trends in computer science were detected and analysed using co-word network and time series analysis with discrete generalised distribution. The comparative analysis of the four open-access databases was presented in the study and the main differences in usage were highlighted.

The importance of open-access resources in communication in science and areas of expertise was shown, and the scientometric analysis presented examples of using datasets. Comparative analysis of keywords from the co-word network using the databases presented the difference between the same concepts in academic and expert areas. The analysis of the *Arxiv* database highlighted the predominance of physics and the connection between computer science and physics.

Keyword and co-word network analysis were carried out for *Arxiv* and *StackExchange* and for specific keywords. Keywords with the highest centrality and betweenness centrality measures reflect the main concepts. The difference in the concept usage in academic and expert contexts was shown with the example of the topic *optimization*.

The methods and tools presented in the paper could be applied to any area of research and other databases, which allows the development of the area through the academic and expertise perspective to be described, and main concepts and the most important topics to be defined.

Acknowledgements

The findings were obtained in cooperation with prof. Dmytro Lande (NTUU KPI), PhD Barbara Żogała-Siudem (SRI PAS), PhD Grzegorz Siudem (WUT), prof. Marek Gągolewski (SRI PAS, WUT).

Iryna Balagura acknowledges support from the British Academy through the Researchers at Risk Fellowships Programme (Grant RaR\100215).

References

- [1] D. Lande, M.Fu, W.Guo, I.Balagura, I. Gorbov, H. Yang, Link prediction of scientific collaboration networks based on information retrieval, *World Wide Web: Internet and Web Information Systems*23(2020) 2239-2257.doi: 10.1007/s11280-019-00768-9.
- [2] Wei-Min Fan, Wei Jeng, Muh-Chyun Tang First published: 06 June 2022 <https://doi.org/10.1002/asi.24688>
- [3] Khajavi R., Arastoopoor S. Semantic domain comparison of research keywords by indicator-based fuzzy distances: A new prospect (2023) *Information Processing and Management*, 60 (5), DOI: 10.1016/j.ipm.2023.103468
- [4] Singh CK, Barne E, Ward R, Tupikina L, Santolini M (2022) Quantifying the rise and fall of scientific fields. *PLoS ONE* 17(6): e0270131.<https://doi.org/10.1371/journal.pone.0270131>
- [5] Gabrielatos, C. (2018). *Keyness Analysis: nature, metrics and techniques*. In Taylor, C. & Marchi, A. (eds.) *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge. 225-258.
- [6] Deng, F. (2020). *Computerized Corpus Keyword Approaches to Evaluation: A Case Study of Evaluative Attitudes of “the Belt and Road” Reports in Mainstream Media of China and America*. *International Journal of Electrical Engineering Education*. <https://doi.org/10.1177/0020720920923303>
- [7] Jihye Park, Hye Jin Lee, Sungzoon Cho, Hot topic detection in central bankers’ speeches, *Expert Systems with Applications*, Volume 230, 2023, 120563, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2023.120563>.
- [8] Shou, Xintian, Wang, Yumeng, Duan, Chenglin, Yuan, Guozhen, Wei, Namin, Yang, Yihan; Hu, Yuanhui *Knowledge Domain and Emerging Trends of Glucagon-Like Peptide 1 Receptor Agonists in Cardiovascular Research: A Bibliometric Analysis* (2023) *Current Problems in Cardiology*, 48 (8), art. no. 101194 DOI: 10.1016/j.cpcardiol.2022.101194
- [9] Masoumi, N., Khajavi, R. A fuzzy classifier for evaluation of research topics by using keyword co-occurrence network and sponsors information. *Scientometrics* 128, 1485–1512 (2023). <https://doi.org/10.1007/s11192-022-04618-w>
- [10] D. Lande, V. Andrushchenko and I. Balagura, "Data Science in Open-Access Research on-Line Resources," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 2018, pp. 17-20, doi: 10.1109/DSMP.2018.8478565.
- [11] AlRyalat, S. A. S., Malkawi, L. W., Momani, S. M. Comparing Bibliometric Analysis Using PubMed, Scopus, and Web of Science Databases. *J. Vis. Exp.* (152), e58494, doi:10.3791/58494 (2019).
- [12] Fiorenzo Franceschini, Domenico Maisano, Luca Mastrogiacomo, Empirical analysis and classification of database errors in Scopus and Web of Science, *Journal of Informetrics*, Volume 10, Issue 4, 2016, Pages 933-953, ISSN 1751-1577, <https://doi.org/10.1016/j.joi.2016.07.003>.
- [13] DBLP, URL <https://dblp.org/> computer science bibliography
- [14] Ariel Rosenfeld. 2023. Is DBLP a Good Computer Science Journals Database? *Computer* 56, 3 (March 2023), 101–108. <https://doi.org/10.1109/MC.2022.3181977>
- [15] David, Dimitri Percia, et al. "Measuring security development in information technologies: A scientometric framework using arXiv e-prints." *Technological Forecasting and Social Change* 188 (2023): 122316.

- [16] Bonilla, Ana Isabel and Hernández, Tony and Gómez, Isabel Scientometric analysis on a sample of scientists in astronomy, particles physics and multidisciplinary physics in arXiv.org (1996-2006),2007, Proceedings of ISSI 2007 - 11th International Conference of the International Society for Scientometrics and Informetrics, 830 – 831.
- [17] Santos, Breno Santana, et al. "COVID-19: A scholarly production dataset report for research analysis." *Data in Brief* 32 (2020): 106178.
- [18] Santos, Breno Santana, et al. "Discovering temporal scientometric knowledge in COVID-19 scholarly production." *Scientometrics* 127.3 (2022): 1609-1642.
- [19] <https://stackoverflow.com/about>
- [20] Lal, Sangeeta, and Rahul Mourya. "For CS Educators, by CS Educators: An Exploratory Analysis of Issues and Recommendations for Online Teaching in Computer Science." *Societies* (Basel, Switzerland) 12.4 (2022): 116.
- [21] Karbasian, Habib, and Aditya Johri. "Insights for Curriculum Development: Identifying Emerging Data Science Topics through Analysis of Q and A Communities." SIGCSE 2020 - Proceedings of the 51st ACM Technical Symposium on Computer Science Education. N.p., 2020. 192–198.
- [22] Lu, Xuecong et al. "The Impact of Linguistic Complexity on Leadership in Online Q&A Communities: Comparing Knowledge Shaping and Knowledge Adding." *Information & Management* 59.6 (2022): 103675.
- [23] Sinha, Priyanka et al. "A Hierarchical Clustering Algorithm for Characterizing Social Media Users." *The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020*. NEW YORK: Assoc Computing Machinery, 2020. 353–362.
- [24] Li, Xuemei, Mike Thelwall, and Kayvan Kousha. "The role of arXiv, RePEc, SSRN and PMC in formal scholarly communication." *Aslib journal of information management* 67.6 (2015): 614-635.
- [25] Andrushchenko, V. B., I. V. Balagura, and D. V. Lande. "Information resources for scientific information access and exchange, identification of scientists-Opportunities, disadvantages, benefits." *CEUR Workshop Proceedings*. 2016.
- [26] Arxiv URL https://arxiv.org/category_taxonomy
- [27] Lande D.V., Kryuchyn A.A., Dobrovska S.V., Balagura I.V. Use of the «Library of Science Metrics» system for conducting science metric research. *Data Rec., Storage & Processing*. 2023. Vol. 25, No. 1. P. 32–42.
- [28] Lande, Dmytro V., and A. A. Snarskii. "Compactified Horizontal Visibility Graph for the Language Network." *arXiv preprint arXiv:1302.4619* (2013).
- [29] Balagura, I., Kadenko, S., Andriichuk, O., & Gorbov, I. (2019). Defining Potential Academic Expert Groups based on Joint Authorship Networks Using Decision Support Tools. In *CEUR Workshop Proceedings*. pp. 222-233.