Article

Identification of Actual Bibliometric/Scientometric Issues Based on 2018-2022 Data from the Lens Platform by Building Key Term Co-occurrence Network

Boris N. Chigarev*

* Oil and Gas Research Institute of the Russian Academy of Sciences, Moscow, Russia, ORCID: 0000-0001-9903-2800, bchigarev@ipng.ru

Abstract: The purpose of this article is to demonstrate the ability of bibliometric data from the Lens platform to identify relevant bibliometric/scientometric issues based on the construction of a network of key terms co-occurrence and their clustering. The advantages of using the Lens platform for bibliometric analysis are briefly demonstrated. Key terms were selected on the basis of n-grams and noun phrases. VOSviewer, Scimago Graphica, and Sifaka text mining application were used as analytical tools. Analysis of the Lens metadata on bibliometrics/scientometrics showed their predominant use in the political, social, and medical fields of research.

Keywords: bibliometrics; scientometrics; the Lens; key terms; VOSviewer; Scimago Graphica

---

Introduction.

This section briefly summarizes the benefits of the Lens platform for bibliometric analysis.

As of this writing, 249,019,130 scientific papers are indexed in the Lens system, while Scopus has 81 million curated papers. Yes, the carefully controlled selection of journals and articles in the abstract database is important for assessing scholarly activity. But for other tasks, such as identifying actual problems within a certain topic, it is important to cover as widely as possible the research in progress. Many high-quality papers are published as preprints, which are not indexed in Scopus or Web of Science; the same situation exists for industry conferences, whose proceedings are not indexed in Scopus and Web of Science, but are of great interest for identifying promising research topics.

The Lens is an open-access metadata aggregator combining three datasets:

- Scholarly Works: metadata of scholarly literature and tools to analyze it.
- Patents: an extensive collection of patent literature with references.
- PatSeq: a search and analysis tool for biological sequences disclosed in the patent literature.

Only metadata of scholarly literature will be used in this article. The interesting feature of the Lens platform - the citation of Scholarly Works by Patents, is beyond the scope of this article.

The entire software behind the Lens platform is open source and free to use:

- PostgreSQL, MySQL and MongoDB databases
- Elasticsearch and Apache Lucene for indexing and full-text search
- Vega-Lite — high-level grammar of interactive graphics

In other words, one can locally reproduce most of the functionality of the Lens for bibliometric research on a specific scientific topic.

The Lens uses bibliometric data from CrossRef, PubMed, CORE, ORCID, Impactstory, Microsoft Academic, and its successor, OpenAlex.

The Lens provides authors with free access to their Researcher Profile, allowing them to correct discrepancies and add to their records.

The Scholar Analysis section, in my opinion, is superior to the Scopus and WoS analysis sections available without a separate subscription.

Additionally, up to 50,000 records in JSON/JSON lines/CSV/RIS/BibTeX formats are available for exporting metadata from the Lens per download. The bibliometric data in JSON lines format is not only more complete than in CSV, but also allows you to use both JSON and string utilities, such as sed, grep, etc.

The charts and data from Scholar Analysis can not only be exported in a variety of formats, but also saved as Dashboards, which can then be made accessible via a link to colleagues.

The system of filters used in compiling queries is very advanced, for example, the Subject Matter filter consists of: Subject, MeSH Heading, Field of Study, Chemical Substance Name, Keyword.

The above capabilities determined the choice of the Lens as a source of bibliometric data of scientific publications on the subject of bibliometrics/scientometrics.

Another important argument was the rare use of the Lens compared to Scopus and Web of Science (about 40 times) in bibliometric studies. Therefore, the motivation arose to show some of the possibilities of the Lens.

To learn more about the Lens platform, it is advisable to read the presentations prepared by its employees [1,2].

Below are several recently published articles illustrating the use of the Lens platform data in scientific articles.

The paper [3] analyzes new sources of citation data, such as Microsoft Academic, Dimensions, and the OpenCitations Index of CrossRef open DOI-to-DOI citations (COCI). The 3,073,351 citations found by these data sources to the 2,515 English-language highly cited documents published in 2006 were analyzed. The authors conclude that Microsoft Academic found more citations in most categories than Scopus and WoS. In many subject categories, Microsoft Academic and Dimensions are good alternatives to Scopus and WoS in terms of coverage.

It is worth clarifying that for the Lens platform, both in 2020 and in 2006, the data from Microsoft Academic composed the majority of materials. In 2022, the Lens began to collaborate with OpenAlex, the successor to Microsoft Academic. As noted above, the Lens also uses data from CrossRef, PubMed, CORE, ORCID, and Impactstory.

Therefore, the article confirms that open-access alternatives to Scopus and WoS can be useful for bibliometric analysis.

The good examples of using the Lens data along with other data can be found in the following publications.

In [4], the authors analyze initiatives that promote local development through the proper use and management of endogenous territorial opportunities to achieve economic, social and environmental development. In order to understand how the field of territorial development has evolved over time, subscription-based data sources (Scopus, Science Direct, Ebsco, and Web of Science) and open access (Lens and Dimension platforms) were used to study research topics and groups of research topics.

The authors [5] conducted a systematic review of articles published between 2015 and 2021 related to academic supervision and collected data by documenting them and reviewing them. A total of 25 articles were collected from Google Scholar, DOAJ, and lens.org. The use of academic supervision by principals has been found to improve teacher performance as well as encourage teacher development.

The authors [6] conducted a systematic search of the literature using PubMed, Embase, Scopus, Web of Science, Science Direct, MedRxiv, and Lens.org databases, which contained studies reporting anaphylaxis after vaccine injection. The conducted systematic review included 41 studies which reported anaphylaxis. A total of 7,942 cases, with 43 deaths, were reported in 14 countries. Most cases occurred after the first vaccine dose was injected. It is important to note that the benefits of vaccination outweigh the risks of anaphylaxis.

The article [7] provides a detailed description of the research analysis process using a systematic literature review. Sixty-nine studies were selected from five different online libraries: ACM, DOAJ, Lens.org, SCOPUS, and SpringerLink. The following conclusions were obtained from the review: sexual and gender diversification education is prevalent in health care, there is a lack of research on the topic in Latin America, and technological tools are minimally used in the teaching process.

It should be noted that none of the aforementioned works used exclusively the Lens data, and that they are more about systematic reviews than bibliometric analysis.

This fact also motivated a more detailed demonstration of the Lens platform's capabilities for at least one type of bibliometric analysis.

Materials and Methods

The data for ongoing bibliometric study were exported from the Lens abstract database using Field of Study filters: Scientometrics, Bibliometrics, and Bibliometric analysis.

Only metadata of the following document types were used: journal article and conference proceedings article.

The time range of the publications was 2018 to 14.09.2022.

The query to the Lens database “Filters: Year Published = (2018 -) Publication Type = (journal article, conference proceedings article) Field of Study = (Scientometrics, Bibliometrics, Bibliometric analysis)” resulted in 10,437 records.

Filtering by Field of Study provides enhanced metadata sampling for publications indexed in the Lens. For example, an article on thematic “trend analysis” may not include the term bibliometric analysis, but the Lens platform is likely to assign it to Field of Study = Bibliometrics OR Bibliometric analysis.

Using only Publication Type = (journal article, conference proceedings article) significantly reduces the sample, otherwise it would contain 13,712 records, but it makes the sample more consistent.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

The Year Published = (2018 -) range gives a sample of records for the last incomplete 5 years, which is sufficient to assess actual issues in publications on Scientometrics OR Bibliometrics.

The choice of analytical tools was made taking into account that the main goal of this publication is to demonstrate the possibilities of using the bibliometric data of the Lens platform to identify actual bibliometric/scientometric issues based on the construction of a key term co-occurrence network and their clustering. The free VOSviewer program, the most widely used for bibliometric analysis, meets these conditions [8].

VOSviewer gives 4,487 Scholarly Works in the fields: Title, Abstract, Full Text. The next popular program CiteSpace — 3,312. VOSviewer utilizes one of the best clustering algorithms and high-quality graphics.

To visualize the slices extracted from the exported by VOSviewer data, a very convenient, yet not often used program Scimago Graphica was used. [9].

The list of noun phrases was obtained using the Sifaka text mining application [10].

Research results.

Actual issues in publications can be defined by a set of the most representative key terms, which not only describe the semantics of the texts under consideration, but also help to choose publications for their further use, for example, for a systematic review.

Unfortunately, the Keywords field in the bibliometric metadata exported by the Lens is incomplete. Thus, in our case 8,112 out of 10,437 records have empty Keywords field.

If the Keywords field cannot be used to create a term co-occurrence map for describing the topics of publications, the Title and Abstract fields can be analyzed to build a term co-occurrence network. This is a built-in procedure in VOSviewer. In this process, it is advisable to remove stop words from the texts and to normalize some terms. Here, we just remove 973 stop words.

This list was assembled from 1,298 stop words posted on GitHub[1], from which terms such as "why'd" have been removed as not normally used in scientific publications, but lengthening the list of words that should be removed from the title and abstract texts.

The procedure for preparing a file to use the Title and Abstract texts to build a term network with VOSviewer was as follows. First, the Title and Abstract texts are combined into one field. This text field is then converted to lower case and the stop words are removed from it with the sed utility. After that, stemming is performed using the Krovetz algorithm[2].

The total number of terms identified by VOSviewer is 124963, of which 1640 terms occur 5 or more times. The 500 terms with the greatest number of links to other terms were used to build the network. To obtain a small number of clusters, the minimum number of terms in a cluster was assumed to be 40. Why 500 terms? A large number of terms do not display well in print materials. For a more detailed examination of the terms network, it is advisable to browse the results in the program itself. With these parameters, the number of clusters obtained was comparable to the other results presented in this paper. The clustering was tested for

[1] github.com/Alir3z4/stop-words/blob/master/english.txt
[2] lexicalresearch.com/software.html

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

stability: the parameter of minimum number of terms in a cluster was varied by 10-20%, while the obtained clustering remained the same.

The results are shown in Figure 1.

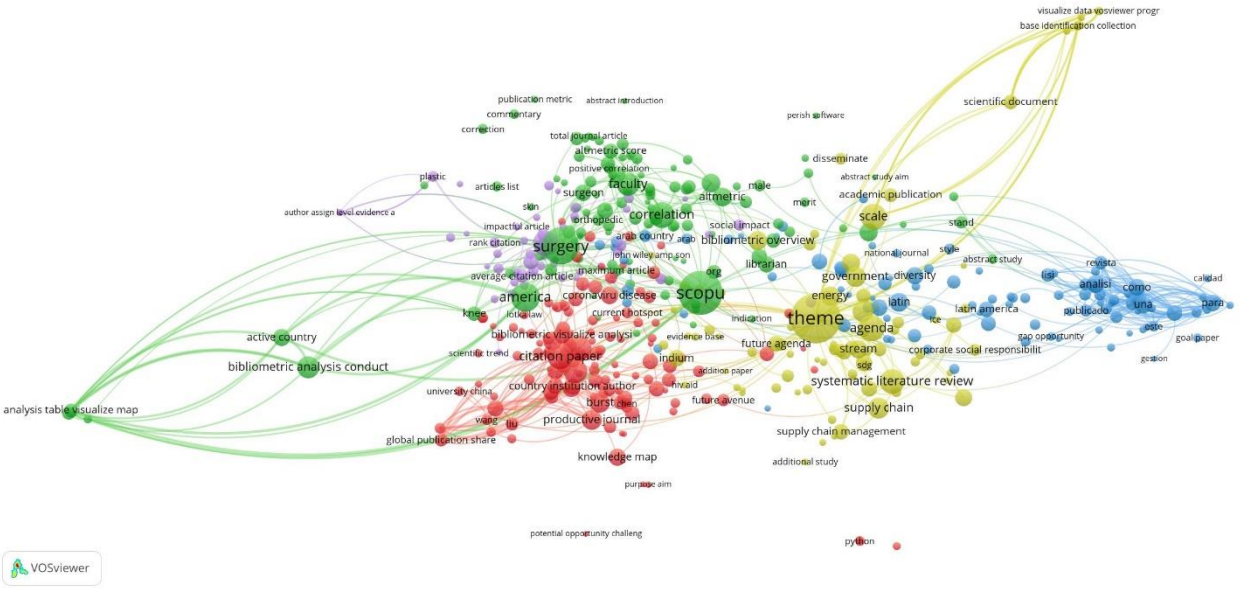


Figure 1. Co-occurrence network of terms obtained by analyzing the pre-processing texts of titles and abstracts of 10,437 bibliometric records.

The list below represents the 10 most common terms for each cluster:

- Cluster 1 (red): citation paper (84); publication output (76); citespace software (75); burst (51); productive journal (51); hotspot field (48); coronavirus disease (41); hot spot (41); bibliometric visualize analysis (40); knowledge map (39).
- Cluster 2 (green): scopus (262); surgery (187); america (112); correlation (89); faculty (67); bibliometric analysis conduct (63); altmetric (48); gender (47); representation (45); active country (40).
- Cluster 3 (cyan): latin (47); analisi (41); una (40); diversity (38); como (37); dissertation (37); para (37); meeting (32); investigacion (31); latin america (31).
- Cluster 4 (khaki): theme (335); agenda (91); scale (90); systematic literature review (80); supply chain (59); entrepreneurship (58); government (55); energy (50); bibliometric overview (46); series (45).
- Cluster 5 (orchid): level evidence (85); citation range (30); list article (26); average citation article (23); original articles review (19); citation classic (18); purpose bibliometric analysis (17); social impact (17); cancer center (16); gamification (16).

The use of title and annotation texts in VOSviewer gives slightly interesting results, and does not reveal a number of program features: (overlays years, aver norm citations).

To describe the subject of publications, you can choose Field of Study, using it similarly as Index Keywords in the bibliometric data of the Scopus platform. VOSviewer itself does not offer this option when loading data exported from the Lens, there are only Author keywords and MeSH keywords, but nothing prevents simply renaming the fields.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

The results of clustering when using the Field of Study field as keywords, are shown in Figure 2.

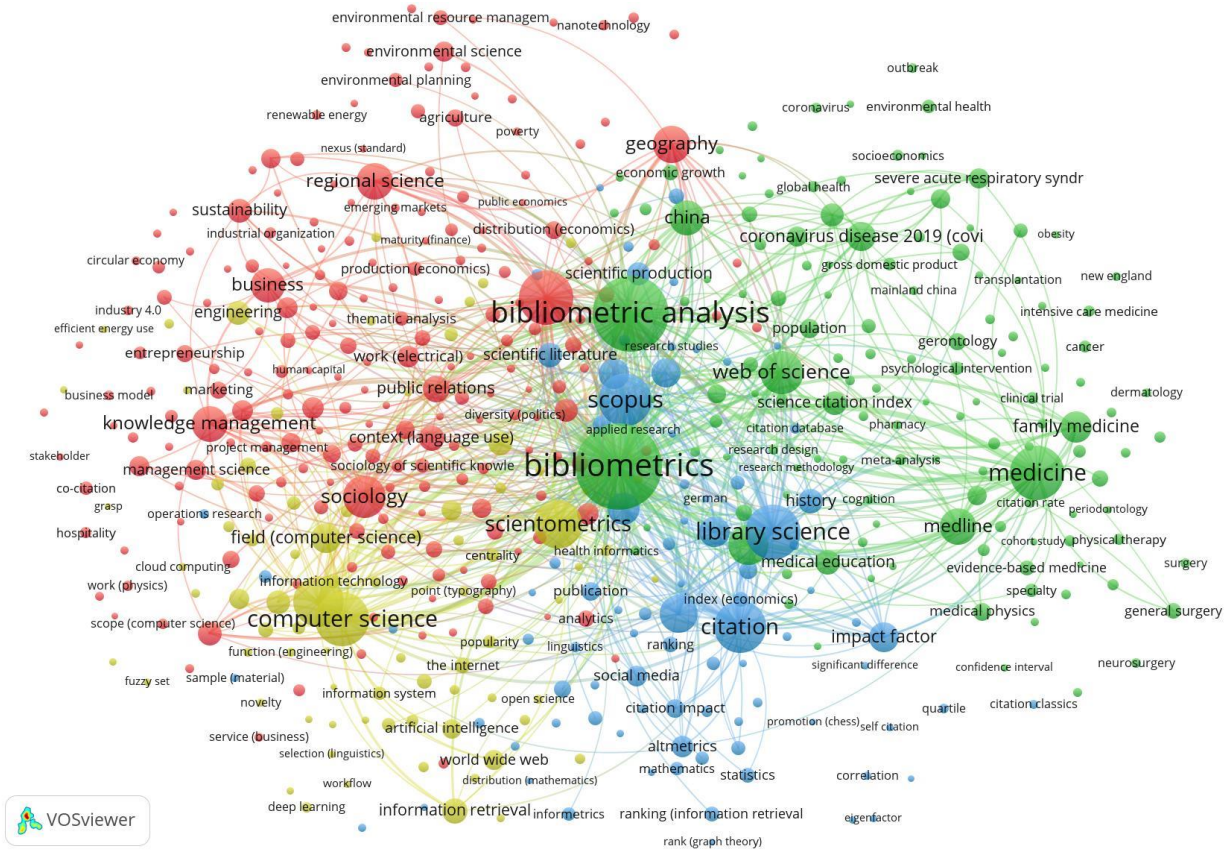


Figure 2. The term co-occurrence network obtained by analyzing the Field of Study as keywords.

If we take not particular clusters, but rather Field of Study terms that occur more than 1,000 times in the total data of 10,437 records, the following list emerges: bibliometrics (5849); bibliometric analysis (4326); library science (2013); computer science (1967); political science (1949); medicine (1930); Scopus (1794); citation (1762); data science (1658); scientometrics (1458); Web of Science (1134); sociology (1089).

The terms given in the list qualitatively represent the general themes of bibliometrics/scientometrics publications. Even the term “library science”, which appears only 8 times in the Title field of 10,437 entries, appears 2013 times in the Field of Study field, as listed above. The reason is that according to the classification given in Scimago Journal & Country Rank, the subject of bibliometrics/scientometrics is in the category Library and Information Sciences, which explains the high value of Field of Study = "library science".

Depending on the purpose of the study, extending the query to the abstract database to include the Field of Study = “library science” filter can yield additional interesting results, for example, concerning the data formats or analytical tools used in the topic “library science”.

According to the list, bibliometric/scientometric methods are most often met in the fields of research: political science (1949); medicine (1930); sociology (1089).

Next, here are lists of the 10 most frequently occurring Field of Study terms for each of the four clusters in Figure 2.

- Cluster 1 (red): political science (1949); sociology (1089); geography (803); regional science (753); knowledge management (729); business (654); public relations (363); context (language use) (317); field (mathematics) (276); sustainability (269).
- Cluster 2 (green): bibliometrics (5849); bibliometric analysis (4326); medicine (1930); web of science (1134); psychology (970); medline (762); china (672); family medicine (515); coronavirus disease 2019 (covid-19) (433); medical education (294).
- Cluster 3 (cyan): library science (2013); scopus (1794); citation (1762); citation analysis (830); subject (documents) (542); social science (501); publishing (484); impact factor (478); productivity (433); scientific literature (294).
- Cluster 4 (khaki): computer science (1967); data science (1658); scientometrics (1458); field (computer science) (463); information retrieval (321); field (geography) (278); engineering (277); visualization (273); world wide web (193); social network analysis (184).

Obtained clusters so well describe individual relevant bibliometric/scientometric problems, that they can be used to conduct independent bibliometric research, applying these terms as filters in the Lens system.

In the list above, the significance of the results was evaluated according to the criterion of occurrence of terms; a similar procedure can be carried out using the average normalized citation indicator. If the first criterion can be defined as "what people most often write about," the second criterion can be defined as "what people most often read about".

A list of 10 Field of Study terms for each cluster, selected according to the criterion of average normalize citation:

- Cluster 1 (red): resource (project management) (6.6854); focus (computing) (6.1986); circular economy (6.0688); greenhouse gas (4.3259); industry 4.0 (3.6947); creativity (3.6658); business model (3.6303); research opportunities (3.1395); nanotechnology (2.9831); human health (2.9242).
- Cluster 2 (green): workforce (2.8769); outbreak (2.8663); scientific publishing (2.4613); infectious disease (medical specialty) (2.2624); gender gap (2.2327); global health (2.1039); pandemic (2.0918); coronavirus (2.0584); knowledge structure (2.0235); inequality (2.0006).
- Cluster 3 (cyan): citation data (3.389); information dissemination (1.8371); regression analysis (1.7685); operations research (1.6657); science policy (1.5303); protocol (science) (1.5256); dimension (data warehouse) (1.5239); bibliographic database (1.4898); scientific progress (1.4623); scientific productivity (1.4512).
- Cluster 4 (khaki): status quo (2.8413); smart city (2.3625); blockchain (2.2695); machine learning (2.2042); data source (2.0185); open science (2.0091); cryptocurrency (1.8547); field (computer science) (1.7902); deep learning (1.7792); term (time) (1.6447).

Data from the list can be used to get a clue as to which problems enumerated for each cluster are likely to attract more attention and citations if analyzed.

A list of 10 Field of Study terms for each cluster, selected according to the average year of publication:

- Cluster 1 (red): field (mathematics) (2021.8043); theme (computing) (2021.7273); work (physics) (2021.6905); process (computing) (2021.6452); quality (philosophy) (2021.6364); government

 doi:10.20944/preprints202212.0533.v1

(linguistics) (2021.6154); scope (computer science) (2021.6053); extant taxon (2021); management science (2020.8377); biodiversity (2020.8182).

- Cluster 2 (green): alternative medicine (2021.1935); dentistry (2021.0435); coronavirus disease 2019 (covid-19) (2020.9769); severe acute respiratory syndrome coronavirus 2 (sars-cov-2) (2020.9686); pandemic (2020.9685); clinical psychology (2020.9565); 2019-20 coronavirus outbreak (2020.9086); traditional medicine (2020.8261); diabetes mellitus (2020.7895); oral and maxillofacial surgery (2020.75).
- Cluster 3 (cyan): rank (graph theory) (2021.7857); context (archaeology) (2021.7164); sample (material) (2021.6957); index (typography) (2021.6375); distribution (mathematics) (2021.625); promotion (chess) (2021.625); operations research (2021.25); grey literature (2020.8235); identification (biology) (2020.7586); econometrics (2020.7333).
- Cluster 4 (khaki): domain (mathematical analysis) (2021.7619); grasp (2021.25); cryptocurrency (2020.7895); efficient energy use (2020.7059); frontier (2020.5909); timeline (2020.5714); data mining (2020.5556); informatics (2020.5455); betweenness centrality (2020.5238); data science (2020.5145).

This list complements the previous two lists with terms indicating which issues are most relevant in recent publications.

When using Field of Study as key terms, the above results are useful for future case studies, when used as filters in constructing subsequent queries to the Lens.

The Field of Study terms are a controlled vocabulary that is composed for the entire corpus of texts in the Lens system. For each topic of study, it is desirable to have a specialized vocabulary compiled from the corpus of texts related to that topic. As a simplified vocabulary, terms obtained by identifying n-grams OR noun phrases in title and abstract texts of bibliometric data exported from the Lens on a given query can be used.

The compilation of the list of n-grams depends significantly on the preliminary preparation of the texts of titles and abstracts.

When querying abstract databases, we are not interested in n-terms like “in this article”, which are often found in abstracts, so stop words need to be removed from the text beforehand. In this section, as above, we used a list of stop words consisting of 973 terms.

Krovetz stemming was used after removing stop words. Stemming is often a bit redundant, but it does not much matter for queries to the abstract database, since the search engines themselves use a similar stemming. Krovetz, without any further adjustments, returns the term Scopus as “scopu”. A direct check confirms that queries to the Lens “scopus” and “scopu” yield the same result. The Lens uses Elasticsearch as its search engine, which in turn uses different stemmings, including Krovetz.

Elasticsearch recommends using porter_stem stemmer for English, but this algorithm tends to stem more aggressively than the kstem (Krovetz), which leads to a more difficulty in human perception of the text once an aggressive stemming is performed. Furthermore, besides algorithms, kstem relies on dictionary, which allows manual correction of stemming for texts of a particular subject area.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

Once the texts of the titles and abstracts had been pre-processed, a list of n-grams was generated. The total number of n-grams occurring 2 or more times in our corpus of texts was 156665. To reduce the vocabulary list of n-grams, we chose an occurrence threshold of 10, which resulted in a vocabulary of 11611 terms. Spaces in n-grams have been replaced with underscores in the resulting dictionary (this is a purely technical convenience, turning n-grams into single words). Next, n-grams were replaced by n-grams with an underscore in the original texts, and the other terms were deleted. The result was a set of terms, which can be considered as a set of keywords for this record, compiled from the n-grams obtained from the title and abstracts texts. The obtained data were entered into the Index Keywords field, and the final table was used in the VOSviewer program for conducting the analysis.

The results of the work done are shown in Figure 3.

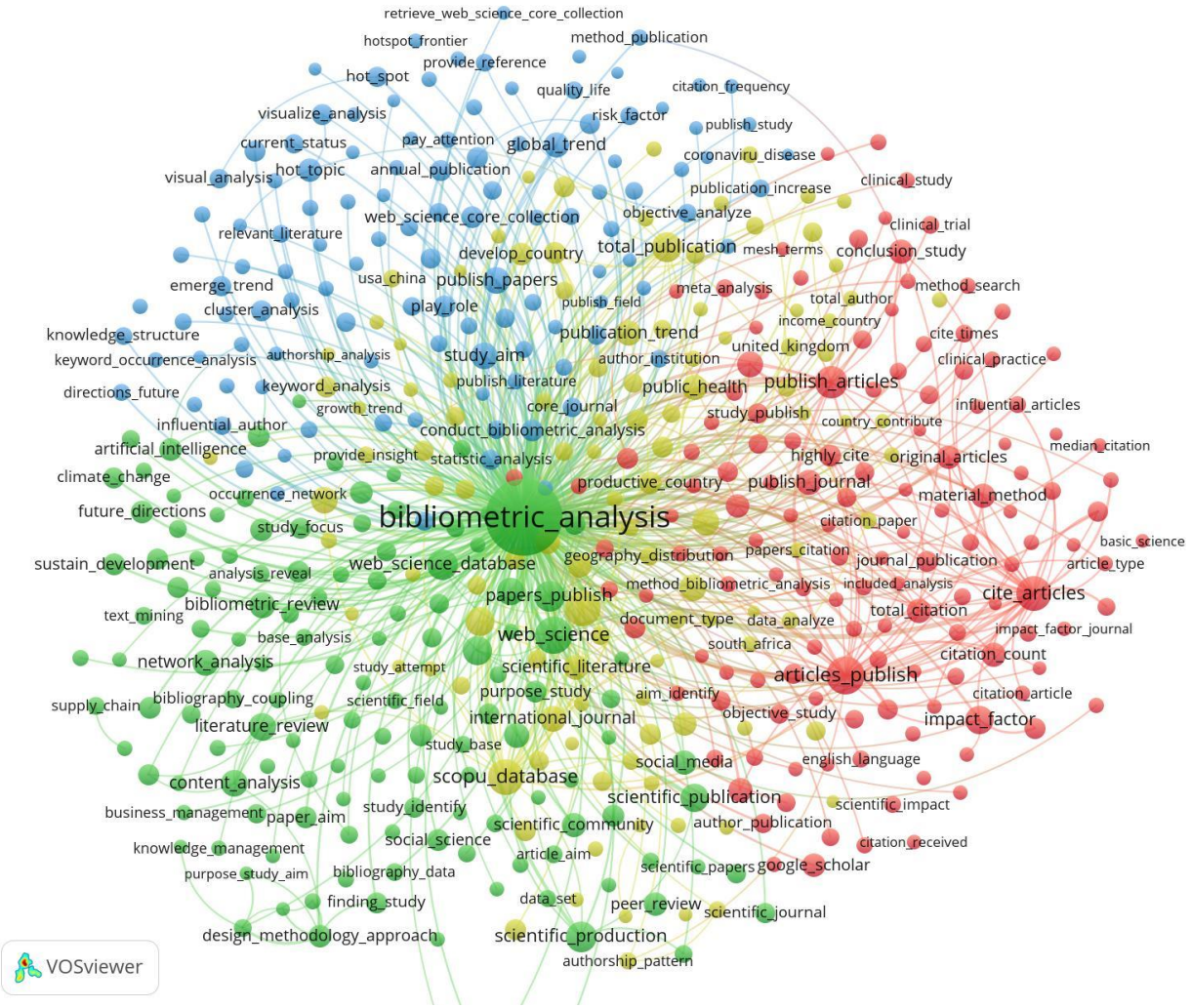


Figure 3. Network of terms co-occurrence obtained by using the n-gram of titles and abstracts texts as keywords.

The following terms generated from the n-gram are the most common in all records:

bibliometric_analysis (2472); web_science (448); scopus_database (410); bibliometric_study (393); cite_articles (370); web_science_database (284); scientific_production (280); total_publication (267); scientometric_analysis (253); citation_analysis (250); impact_factor (226).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

Some frequent phrases like articles_publish have been removed from this list. The term bibliometric_analysis significantly dominates the presented list; hence it is reasonable to use it in queries in future studies.

By clusters, the most frequent terms have the following distribution:

- Cluster 1 (red): articles_publish (465); cite_articles (370); publish_articles (324); impact_factor (226); total_articles (195); publish_journal (178); journal_articles (168); material_method (167); citation_count (162); total_citation (161). In this cluster, the articles are the basis for the formation of bibliometric data.
- Cluster 2 (green): bibliometric_analysis (2472); web_science (448) + web_science_database (284); papers_publish (282); scientific_production (280); citation_analysis (250); scientific_publication (228); systematic_review (221); literature_review (220); content_analysis (205); network_analysis (189); bibliometric_review (183). Here the emphasis is on the types of analysis and the types of publications in which they are used.
- Cluster 3 (cyan): study_aim (199); global_trend (181); hot_topic (152); conduct_bibliometric_analysis (129); play_role (129); development_trend (117); visualize_analysis (116); cluster_analysis (114); country_institution (114); annual_publication (113); emerge_trend (109); trend_field (107); hot_spot (106). In this list, some general terms have been removed to emphasize the importance of trend detection: hot_topic, global_trend, emerge_trend, and the relevance of visualize_analysis and cluster_analysis in conduct_bibliometric_analysis.
- Cluster 4 (khaki): scopus_database (410); bibliometric_study (393); total_publication (267); scientometric_analysis (253); international_collaboration (206); vosviewer_software (203); covid_pandemic (197); international_journal (190); scientific_literature (181); productive_author (171). In my opinion, the most interesting terms in this list are international_collaboration (206) and vosviewer_software (203).

In lists above are only the terms selected by the highest frequency criterion; the other samples, unlike the Field of Study data, will be presented graphically.

Modern tools of visual analysis (see cluster 3), allow to present such data slices in the form of vivid graphs. For this purpose, it is very convenient to use the possibilities of Scimago Graphica free program, a detailed description of which is given in [9].

Data slices were generated from files exported from VOSviewer. Bubble plots allow four variables to be displayed in a single figure: the two axes, the size of the bubble, and its color.

In Figure 4, the average normalized citations of papers containing a particular term and the average year of such publications are plotted along the axes, the frequency of occurrence of the term is represented by the size of the bubble, and the cluster to which the term belongs is indicated by color.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

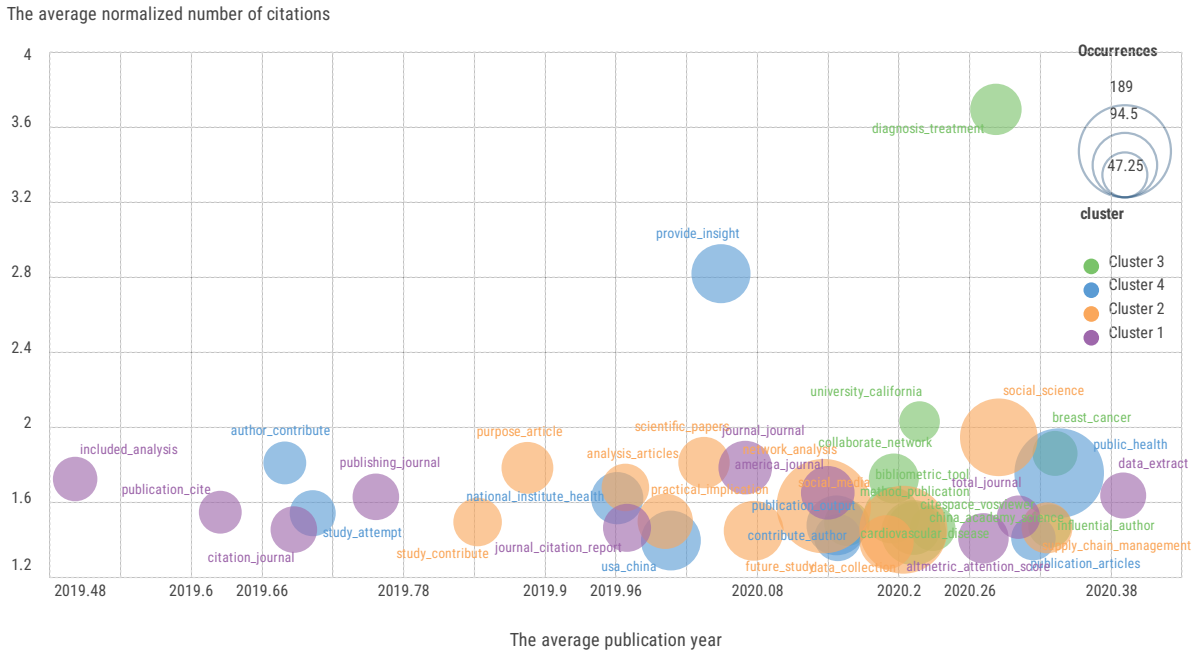


Figure 4. Bubble chart of 40 terms with the highest average normalized citation vs. average year of publication in four clusters.

Diagnosis treatment, provide insight, public health, breast cancer, social science, and data extract – these terms belong to different clusters, occur in the second half of the time period studied, and reflect a promising, socially relevant topic of data extraction for diagnosing important public health problems such as breast cancer.

Using the Bubble chart offers great flexibility in how to display even the same data. For example, in Figure 5, the emphasis is on the list of terms and their average normalized citation. The average year of publication is auxiliary.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

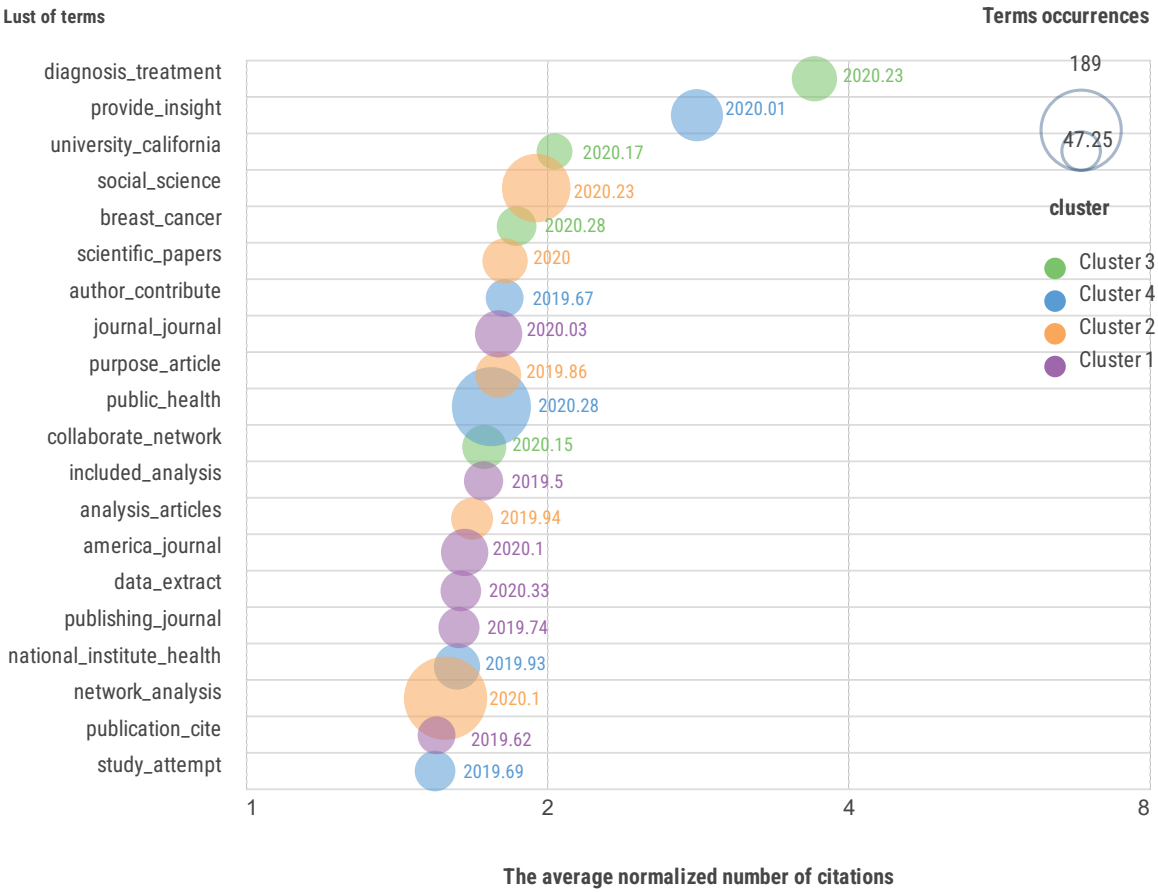
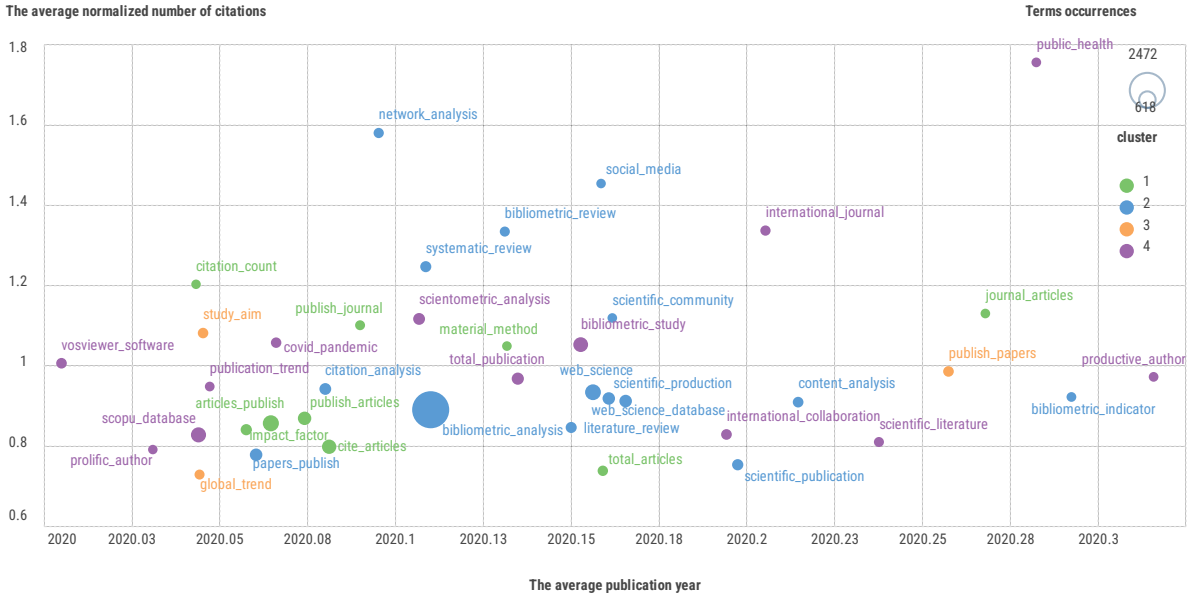


Figure 5. Bubble chart of 40 terms with emphasis on the list of terms and their average normalized citation.

Another informative approach in a bubble chart visual analysis of the data is to change the sample that will appear in the graph.

If, in the previous example, the sample consisted of the 40 terms with the highest average normalized citation, then the chart #6 is plotted for the sample of terms most frequently occurring in titles and abstracts. It is clearly visible that in this case the scale of the "Occurrence" legend is much larger than in the previous two graphs.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

Figure 6. Bubble chart of 40 terms with the highest frequency of occurrence in coordinates of average normalized citation and average year of publication for four clusters

The graph shows that both the set of terms themselves and the topics they describe have changed, which relate more to the bibliometric analysis itself than to the promising topic for analysis, as it was in the previous case.

The global picture of the network of terms co-occurrence obtained by VOSviewer, shown in Figure 7, can be viewed in more detail in the program itself by changing the scale of the image. However, the data given in the publication do not have such an opportunity. Therefore, the Scimago Graphica program can be used to display the desired slice of the data exported from VOSviewer. An example sampling for 30 terms from the most cited publications is shown in Figure 7 as a slice of total term co-occurrence network.

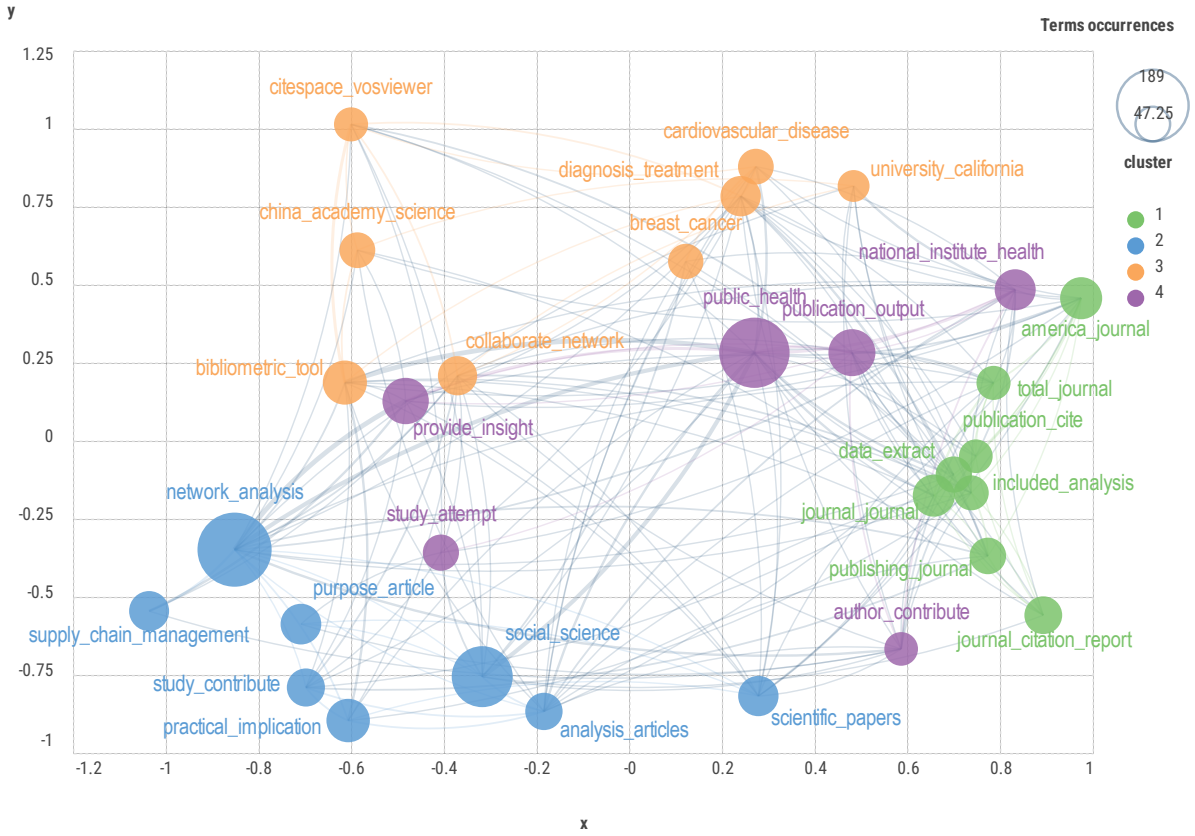


Figure 7. A slice of total term co-occurrence network for 30 terms with the highest average normalized citation.

This figure used coordinates taken from data exported from VOSviewer. In the graph, links between different clusters are marked in gray, and those belonging to the same cluster are marked in cluster colors. The figure shows that the terms of the third cluster are more linked than those of the other clusters, which means that they can be found together more often in the articles on the topic of the third cluster.

As for the Bubble chart, various data slices can be prepared for the term co-occurrence network in order to analyze it in more detail. While the previous graph shows terms from publications with the highest average normalized citation, Figure 8 shows the network of the 34 most frequent terms.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

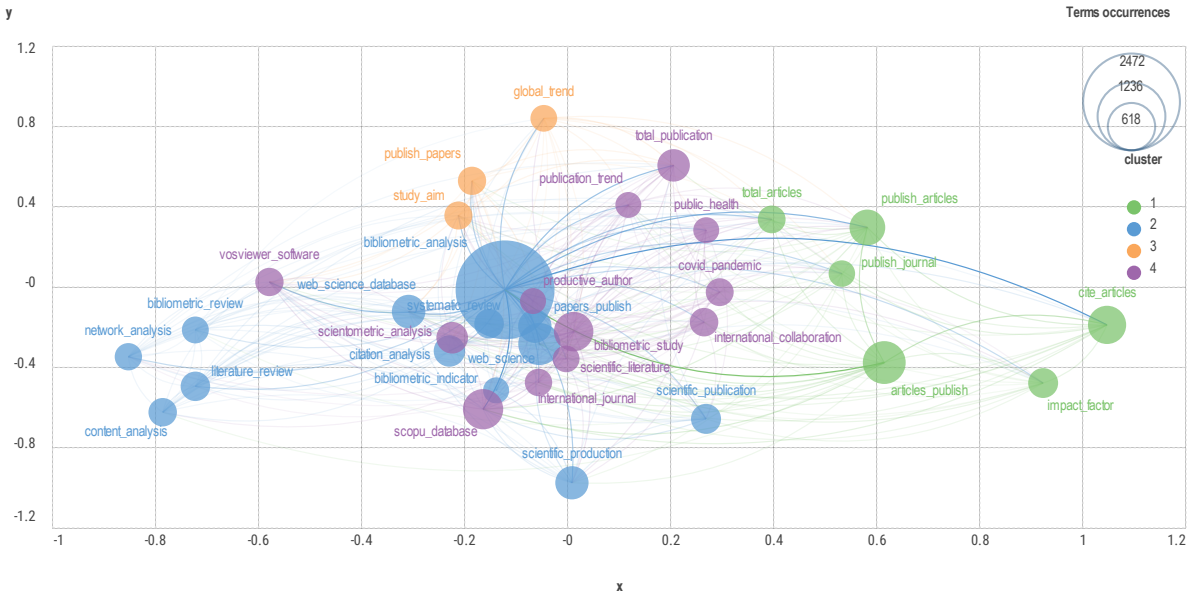


Figure 8. A network of the 34 most common terms. Taken from data exported from VOSviewer.

The graph shows that the term bibliometric_analysis is the dominant one, and the other most frequently encountered terms describe well the main aspects of the cluster topic.

If taking the average year of publication and the average normalized citation, rather than the old coordinates taken from exported data from VOSviewer, the picture comes out as follows:

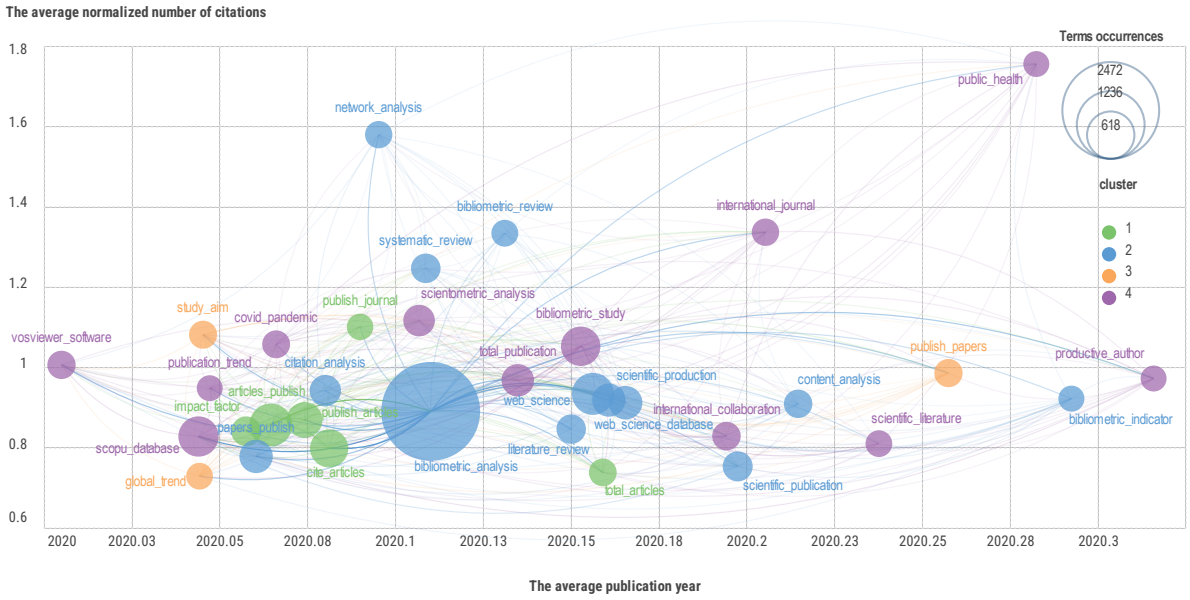


Figure 9. A network of the 34 most common terms. Using the average normalized citation and the average year of publication as coordinates.

This graph is more informative than the previous one, because it reflects more meaningful variables. Therefore, it is advisable to use it. The graph actually repeats figure 6, but it also shows the relationships between the terms, which makes it more informative. In fact, to the four bubble-plot variables, a fifth is added — the relationship between the terms (nodes).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0533.v1

Final remarks. Using Scimago Graphica program allows conducting a graphical analysis of different slices of the global network of terms co-occurrence and a detailed examination of the significance of the constraints under which the samples were taken for the visual analysis.

N-grams are only one form of meaningful collocations in a text. Another form is noun phrases. In order to show the significance of these collocations, the Sifaka program was used in this work. As noun phrases were formed using morphological tagging, stemming and removing word stops were not done. A separate file was generated from the titles and abstracts of each bibliometric record to form a corpus of texts for further indexing. Indexing was performed using sifakaBuildIndex.jar, and a list of the 2000 most frequent noun phrases was generated using sifakaTextMiner.jar.

The following is a short list of noun phrases and their occurrence in the resulting corpus of texts in the format: Term (Document Frequency) (Collection Term Frequency):

scopus database (827) (972); research trends (723) (880); science database (630) (671); h index (376) (617); research hotspots (404) (546); literature review (377) (497); research topics (320) (395); science core collection (358) (390); research output (296) (390); research articles (296) (362); research field (276) (349); network analysis (273) (322); research directions (283) (313); vosviewer software (296) (303); covid pandemic (192) (295); science citation index (246) (282); impact factor (209) (269); citation analysis (222) (255); journal articles (208) (249); research productivity (176) (248); research areas (221) (247); science core collection database (228) (232); content analysis (186) (226); originality value (220) (220); research status (184) (216); research area (185) (207); publication trends (171) (199); citation count (143) (198); google scholar (149) (197); research papers (166) (196); research publications (157) (193); information science (121) (170); growth rate (137) (167); co-occurrence (154) (163); citation impact (119) (163); year period (138) (154); citation counts (114) (143); research fields (118) (143); research agenda (119) (142); case study (121) (140); research topic (125) (137); design methodology approach (135) (136); climate change (77) (136); research themes (113) (133); research frontiers (109) (132); bibliometrics analysis (117) (131); co citation analysis (115) (130); co authorship (111) (129); research institutions (110) (124); computer science (107) (122).

The presented noun phrases are very meaningful, so it is reasonable to use them as key phrases when making queries to abstract databases.

Removing stop words from our corpus of texts significantly changes the list of frequently occurring noun phrases and reduces their total number. This illustrates the advisability of making minimal changes to the source texts before performing morphological analysis to detect noun phrases:

web science (410) (441); scopus database (394) (423); literature review (237) (302); web science database (267) (274); purpose study (248) (248); aim study (222) (222); science citation (159) (169); network analysis (143) (166); web science core collection (157) (164); analysis publications (140) (148); purpose paper (133) (134); vosviewer software (132) (133); journal articles (119) (132); citation analysis (116) (129); analysis articles (122) (124); web science core collection database (121) (121); growth rate (94) (107); content analysis (91) (104); materials methods (99) (99); citations paper (70) (87); publication trends (81) (85); design methodology approach (82) (82); publications citations (79) (81); analysis literature (77)

 doi:10.20944/preprints202212.0533.v1

(79); data analysis (72) (75); income countries (53) (70); trends field (69) (69); citation count (64) (69); publications field (62) (63); academy sciences (61) (63); conclusion study (62) (62); impact factor (57) (62); study aims (60) (60); conclusions study (60) (60); google scholar (47) (59); vos viewer (57) (59); review literature (55) (58); citations article (55) (57); majority articles (54) (56); covid pandemic (49) (55); keyword analysis (53) (55); analysis papers (55) (55); originality study (55) (55); studies field (53) (54); increase publications (51) (53).

Note: terms are presented as they are received when exported from the program, without spelling corrections.

The identified noun phrases can serve as a good addition to the list of n-grams.

To confirm this, let us compare the lists of 2,000 name phrases and the 1,11611 n-grams derived from the same texts. Despite the huge number of n-grams, only 785 name phrases have the similar meaning with them. Examples of non-matching terms in the lists: research topics, research output, research area, research fields, co-occurrence, co citation, visualization analysis, decision making, research landscape, regression analysis. Some of the mismatch is caused by stemming, such as: research topics vs research topic or co-occurrence vs cooccurrence, others are not on the list of n-grams, but are relevant when composing prospective queries: decision making, regression analysis.

The above results demonstrate the feasibility of creating a controlled thematic vocabulary and bringing the terms to a normalized form to identify the actual issues of a particular topic by bibliometric analysis based on the co-occurrence of the terms. That is, to perform lemmatization of texts on the basis of controlled thematic vocabulary.

Conclusions

The expediency of using the Lens abstract database for bibliometric analysis has been shown. Its main advantages are: openness, large volume of bibliometric data, possibility to export up to 50 thousand records simultaneously, high-quality graphics of the analytics section, cooperation with OpenAlex, the successor of Microsoft Academy, which has an advanced API to query its database and even download native metadata of publications.

This article considered only one aspect of bibliometrics: identifying the topics of publications by terms co-occurrence, so further conclusions are limited to this task.

The Field of Study field allows qualitative filtering of data when querying the Lens and expanding the selections by identifying the co-occurrence of Field of Study terms during their clustering.

Partial filling of the author's keyword field can be compensated by compiling key terms based on the simplest analysis of the texts of titles and abstracts, the fields of which are well filled in the metadata of publications exported by the Lens.

It is shown that the joint definition of n-grams and name phrases can be useful in composing the key terms to describe the topics under study.

Field of Study terms, n-grams and noun phrases can be used by VOSviewer as keywords in constructing a co-occurrence network.

Exported from VOSviewer data of the network of terms co-occurrences can be further used to visually analyze slices of the obtained data using the free program Scimago Graphica.

The analysis of the collected metadata on bibliometrics/scientometrics showed their wide use in the political, social, and medical fields of research. This is probably due to the demand and widespread use of analytical reports and systematic reviews in these fields of science.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Ballagh, A. Lens.org: A free and open platform for science and technology mapping. [presentation on the Internet]. 2019 Oct 28, [cited 2022 Oct 12]; [20 p.]. Available from: https://research.qut.edu.au/best/wp-content/uploads/sites/244/2019/11/The-Lens-KCI-Workshop-Presentation.pdf
2. Jefferson, O.A. An open platform for discovery, analytics, mapping and management of research works and innovation pathways. [presentation on the Internet]. 2020 Sept 8, [cited 2022 Oct 12]; [20 p.]. Available from: https://about.lens.org/wp-content/uploads/2017/08/OpenConV2-September-20200908-1.pdf
3. Martín-Martín A, Thelwall M, Orduna-Malea E, Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics 2021;126:871–906. doi: https://doi.org/10.1007/s11192-020-03690-4.
4. Galeano-Barrera CJ, Arango Ospina ME, Mendoza García EM, Rico-Bautista D, Romero-Riaño E. Exploring the Evolution of the Topics and Research Fields of Territorial Development from a Comprehensive Bibliometric Analysis. Sustainability 2022;14:6515. doi: https://doi.org/10.3390/su14116515.
5. Imamah N, Churrahman T. Academic Supervision by School Principals for Improving Teacher Performance. KSS 2022:60–9. https://doi.org/10.18502/kss.v7i10.11209.
6. Paul P, Janjua E, AlSubaie M, Ramadorai V, Mushannen B, Vattoth AL, et al. Anaphylaxis and Related Events Following COVID-19 Vaccination: A Systematic Review. The Journal of Clinical Pharma 2022;62:1335–49. doi: https://doi.org/10.1002/jcph.2120.
7. López-Orozco CF, López-Caudana EO, Ponce P. A systematic mapping literature review of education around sexual and gender diversities. Front Sociol 2022;7:946683. doi: https://doi.org/10.3389/fsoc.2022.946683.
8. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010;84:523–38. doi: https://doi.org/10.1007/s11192-009-0146-3.
9. Hassan-Montero Y, De-Moya-Anegón F, Guerrero-Bote VP. SCImago Graphica: a new tool for exploring and visually communicating data. EPI 2022:e310502. doi: https://doi.org/10.3145/epi.2022.sep.02.

 doi:10.20944/preprints202212.0533.v1

10. VandenBerg C, Callan J. Sifaka: Text Mining Above a Search API 2018. Oct 5, [cited 2022 Oct 12]; [5 p.]. Available from: https://arxiv.org/pdf/1810.02907.