
Demonstration the Importance of Pre-processing the Text Fields of Bibliometric Records to Identify Promising Research Tasks. Case Study of Scopus Data on Petroleum Reservoir Engineering

[Boris Chigarev](#) *

Posted Date: 6 March 2025

doi: 10.20944/preprints202503.0387.v1

Keywords: terms co-occurrence; text preprocessing; abbreviations; promising research tasks; VOSviewer; Scimago Graphica



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Demonstration the Importance of Pre-processing the Text Fields of Bibliometric Records to Identify Promising Research Tasks. Case Study of Scopus Data on Petroleum Reservoir Engineering

Boris Chigarev

Oil and Gas Research Institute, Russian Academy of Sciences, Moscow, Russia; bchigarev@ipng.ru

Abstract. *Background.* Nowadays, bibliometric analyses of data from abstract databases are often used to identify relevant research problems in order to rationalize the use of financial and other resources. *The aim of this paper* was to demonstrate the importance of pre-processing the text fields of bibliometric records to construct a term co-occurrence network and the feasibility of subsequently using Scimago Graphica to examine different slices of clustering results in detail in order to identify relevant research topics. *Materials and Methods.* A total of 8051 records exported from Scopus matching a filter (LIMIT-TO (EXACTKEYWORD, 'Petroleum Reservoir Engineering')) over the last ten years were used. VOSviewer and Scimago Graphica were applied for bibliometric analysis. *The results of the study* showed the relevance of using the filter 'LIMIT-TO EXACTKEYWORD' in the query to Scopus; the expediency of disclosing abbreviations in the text fields of records and preliminary clarification of texts; the effectiveness of using filters in the Scimago Graphica program to build a network of co-currency of terms in order to identify promising research topics; the proposal of promising research objectives arising from the analysis, which can be described by the following terms: 1. nanopores, shale oil, pore size, molecular; 2. nanoparticles; 2. It is observed that in some cases terms occurring in the same cluster are not the best choice for querying in order to expand the collection of publications on a given topic. Therefore, it is proposed to conduct a separate study using Apriori class algorithms for this purpose.

Keywords: terms co-occurrence; text preprocessing; abbreviations; promising research tasks; VOSviewer; Scimago Graphica

Introduction

Assessment of the significance of tasks within a given theme can contribute to the rational use of financial, human and other resources necessary for scientific research. Bibliometric analysis of data from abstract databases can serve as a basis for such assessments.

As an example of publications that reflect the above, the following works can be cited.

The paper [1] argues that peer review decisions provide >95% of the funding for academic medical research, so it is important to understand the effectiveness of peer review and how it can be improved.

The allocation of research funding relies on peer review, which can be biased. A study by the authors [2] analyzes interventions in peer review and decision-making to improve research funding practices.

Recently, funding agencies have begun to call for more research into improving funding allocation processes and seeking effective mechanisms for allocating research funds [3].

Currently, VOSviewer is the most widely used program for bibliometric research. [4].

For example, according to a query to The Lens open abstracts database, for the period 2020 to January 4, 2025, the number of scholarly works in which the term VOSviewer appears in the Title, Abstract, Keyword, or Field of Study fields was 15526, CiteSpace — 8105, and Bibliometrix — 3122.

However, the number of publications dealing with pre-processing of text fields, such as Author Keywords or Annotations, is quite rare.

For example, if we add the term “lemmatization” to the previous query about VOSviewer, we find only one book chapter [5] and one article [6], in which the authors analyze bibliometric data from Scopus and Web of Science on digital transformation, cleaned by lemmatization and stemming, and two preprints by the author of this paper [7,8].

And by adding the term “abbreviation”, we also managed to find a book chapter [9] and the author's preprint [10].

This is not to say that a more detailed search in more abstract databases will not find such publications, but only to point out that the use of VOSviewer in combination with text field preprocessing is rare and can be considered an underappreciated but important task. On the other hand, the pre-processing of texts is a classic task in text analysis [11].

The above has defined the objectives of this study related to using VOSviewer to perform bibliometric analysis:

- Analyze some features of the fields “Author keywords” and “Index keywords”, which are valuable for clustering of terms based on their co-occurrence.
- Show the importance of disclosing abbreviations in text preparation before keyword clustering.
- Provide an example of the usefulness and benefits of using abstract texts to identify topical issues through the case study of Petroleum Reservoir Engineering research in comparison to the Author keywords.
- Demonstrate the possibilities of using the Scimago Graphica program for detailed examination in the coordinates “average time of publication” - “average citation” of different slices of clustering results in order to identify relevant research topics.

In this paper, bibliometric data are sequentially analyzed using VOSviewer and Scimago Graphica to identify relevant research problems. No publications by other authors were found in which the parameters obtained from VOSviewer were used to filter the data and construct coordinate axes in Scimago Graphica program.

The text of the article consistently reveals the significance of the listed tasks.

Materials and Methods

The Scopus data (8051 records) corresponding to the query : (SUBJAREA(ENER) AND PUBYEAR > 2013 AND PUBYEAR < 2025 AND (LIMIT-TO (EXACTKEYWORD,“Petroleum Reservoir Engineering”)) AND (LIMIT-TO (DOCTYPE,“ar”)) AND (LIMIT-TO (LANGUAGE,“English”))) current as of November 16, 2024, was used as a basis for the analysis to identify relevant research challenges and to demonstrate some features of the bibliometric records that are useful to consider in doing so.

Key characteristics of the sample

If we use the field “Author Keywords” for term clustering, it should be noted that in the sample under consideration this field is not filled in 1470 cases out of a total of 8051 records.

At the same time, all records of the “Indexed keywords” field are filled. And each record is assigned from 20 to 30 tags.

In the query used, the term “Petroleum Reservoir Engineering” was specified in the Keywords field. A direct search in the data exported from Scopus showed that this term occurs in all entries of the Index Keywords field, once in the Author Keywords field and 5 times in the Abstract field.

Explanation: according to Scopus support hub¹ — “Indexed keywords are chosen by Scopus and are standardized to vocabularies derived from thesauri that Elsevier owns or licenses. Unlike Author keywords, Indexed keywords take into account synonyms, various spellings, and plurals.”

“Standardized vocabularies” actually determines what labels Scopus assigns to a given bibliometric record in the Index Keywords field.

Thus, by using LIMIT-TO (EXACTKEYWORD, “Petroleum Reservoir Engineering”) in the query, we significantly expand the output of records classified by Scopus on the topic of interest. This is important when there is an interest in enhancing the context of the topic of interest, allowing us to analyze what knowledge can be used to extend the topic under study. A promising topic should have a broad research area that can be exploited as a resource.

However, while expanding the sample, it is desirable to ensure that it remains relevant to the topic under study, so below are some characteristics of the sample used in the paper to demonstrate its appropriateness to the stated topic “Petroleum Reservoir Engineering”.

The most frequent values of the SUBJECT AREA field from the file Scopus_exported_refine_values are the following: Energy (8051), Earth and Planetary Sciences (4475), Chemical Engineering (1854), Engineering (1375), Chemistry (1043), Mathematics (779), Environmental Science (463).

The most frequent values of the SOURCE TITLE field from the file Scopus_exported_refine_values are the following: Journal Of Petroleum Science And Engineering (1627), Energy And Fuels (669), Fuel (588), Energies (581), Journal Of Natural Gas Science And Engineering (475), Journal Of Petroleum Exploration And Production Technology (358), Petroleum Science And Technology (283), Geoenery Science And Engineering (254), SPE Journal (238), Energy (173), International Journal Of Greenhouse Gas Control (150), Energy Sources Part A Recovery Utilization And Environmental Effects (148).

The most frequent values of the KEYWORD field from the file Scopus_exported_refine_values are the following: Petroleum Reservoir Engineering (8051), Petroleum Reservoirs (3941), Hydrocarbon Reservoir (1965), Oil Well Flooding (1588), Permeability (1536), Fracture (1485), Enhanced Recovery (1379), Carbon Dioxide (1266), Crude Oil (1252), Enhanced Oil Recovery (1237), Gases (1180), Low Permeability Reservoirs (1139), Porosity (1084), Reservoirs (water) (1070).

The above field values are in good agreement with the Petroleum Reservoir Engineering theme. This indicates the adequacy of the sample used for the analysis.

Although the term “Petroleum Reservoir Engineering” rarely appears in the text fields of the articles themselves (Title, Abstract, Author Keywords), the tags (Index Keywords) assigned to the articles by Scopus can be used to create queries to the database.

The most frequent values of the COUNTRY field from the file Scopus_exported_refine_values are the following: China (4359), United States (1496), Iran (754), Canada (593), United Kingdom (358), Australia (348), Saudi Arabia (226), Russian Federation (191), Brazil (169), India (150), Norway (131), France (121), Germany (117).

The publications of Russian researchers are underrepresented in international journals; for example, Iran has been under sanctions for a long time, but there are much more publications there. Yes, Russia has many high-quality journals on oil and gas, but English-language publications also fulfill a marketing role to promote the interests of research groups. For example, China, with 4359 publications, is comparable to the sum of the publications of the other countries in this list — 4654.

To participate in international projects, not only scientific but also industrial, it is crucial to promote their technologies and competencies and to establish contacts. The identification of technological capabilities of advanced competencies is largely based on the analysis of scientific literature [12], which determines the importance of publications in international journals.

¹ https://service.elsevier.com/app/answers/detail/a_id/21730/supporthub/scopus/kw/index+keywords+field/

Scientific publications play a vital role in promoting and ensuring energy security by providing critical insights, frameworks and data that inform policy decisions [13].

This study is limited by the use of author keywords and text annotations to identify relevant research tasks related to the topic “Petroleum Reservoir Engineering” and the importance of pre-processing the text fields.

Methods and Programs in Use

The following programs were used for bibliometric analysis: VOSviewer [4] and Scimago Graphica [14].

In preparing the data for analysis, a list of abbreviations occurring in the author's keywords was compiled. The 80 most frequent abbreviations were expanded to their full names.

When creating a list of abbreviations, there may be ambiguous abbreviations that were not considered in our case, e.g. (ES) → effective stress, effective simulation.

Identified author keywords differing only in plural or singular are reduced to singular. Keywords distinguished by the presence or absence of a hyphen or short dash were reduced to the form containing only spaces. Some spellings have been replaced, such as huff 'n' puff to huff and puff. Bracketed terms, including markup tags, have been removed. All terms have been lower-cased. For brevity, we will use the term “normalization” to refer to such preprocessing of terms.

The importance of normalization can be seen by the fact that the total number of unique author keywords reduced to lowercase only was 16102 and 15215 after the normalization procedure.

Examples of meaningful substitutions are the abbreviations EOR → enhanced oil recovery, CO₂ → carbon dioxide, CBM → coalbed methane, the plural reservoirs → reservoir, storages → storage, a term containing a hyphen two-phase flow → two phase flow, a keyword containing an abbreviation in parentheses enhanced oil recovery (EOR) → enhanced oil recovery.

Without normalization the annotation texts and taking all 16102 unique author keywords in their original spelling, only 9502 of them occur in 8051 annotation texts ($9502 \cdot 100 / 16102 = 59\%$).

This means that only 59% of the author keywords occur at least once in the annotation texts in their original spelling.

This can be an issue when selecting keywords to build queries for literature searches — experts select the articles they are looking for by their titles and abstracts. Also, the author keyword field may be less populated than the annotation field in abstract databases.

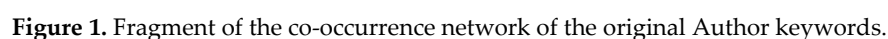
Indexical keywords are not used in this paper, but for reference we note that despite the fact that there are more such words in each entry than author keywords and all entries are populated, unlike author keywords, their occurrence in the annotation texts is somewhat lower, amounting to about 51%.

When constructing the overall landscape of term co-occurrence using the VOSviewer program, the difference between the network of key terms in the original spelling and the normalized one is visually not very striking and requires a more detailed examination of the json files (available in the archive) at app.vosviewer.com. When analyzing particular network slices with the Scimago Graphica program, the difference is more noticeable.

Note: Supplementary materials to the article in the form of graphs, interactive web pages and json data files (for viewing in <https://app.vosviewer.com/>) are available at <https://doi.org/10.6084/m9.figshare.28524263.v1>. These materials allow you to study the graphs in more detail. The file prefixes correspond to the number of the figure in the text of this paper. (Fig_1_name, Fig_2_name, etc.)

Results and Discussions

Comparison of Author Keywords Clustering with and Without Term Normalization



For detailed review of this network, the file AuKWsrGrt4Top400Min50.json placed in the archive can be used at <https://app.vosviewer.com/>.

The network of co-occurrence of author keywords during normalization of terms of the “Author Keywords” field is constructed under the following parameters: the total number of normalized author keywords defined by the program is 15215, of which 969 occur 5 or more times. Using 400 keywords with the greatest total link strength and Minimum cluster size = 50, 5 clusters were obtained, a fragment of which is shown in Fig. 2.

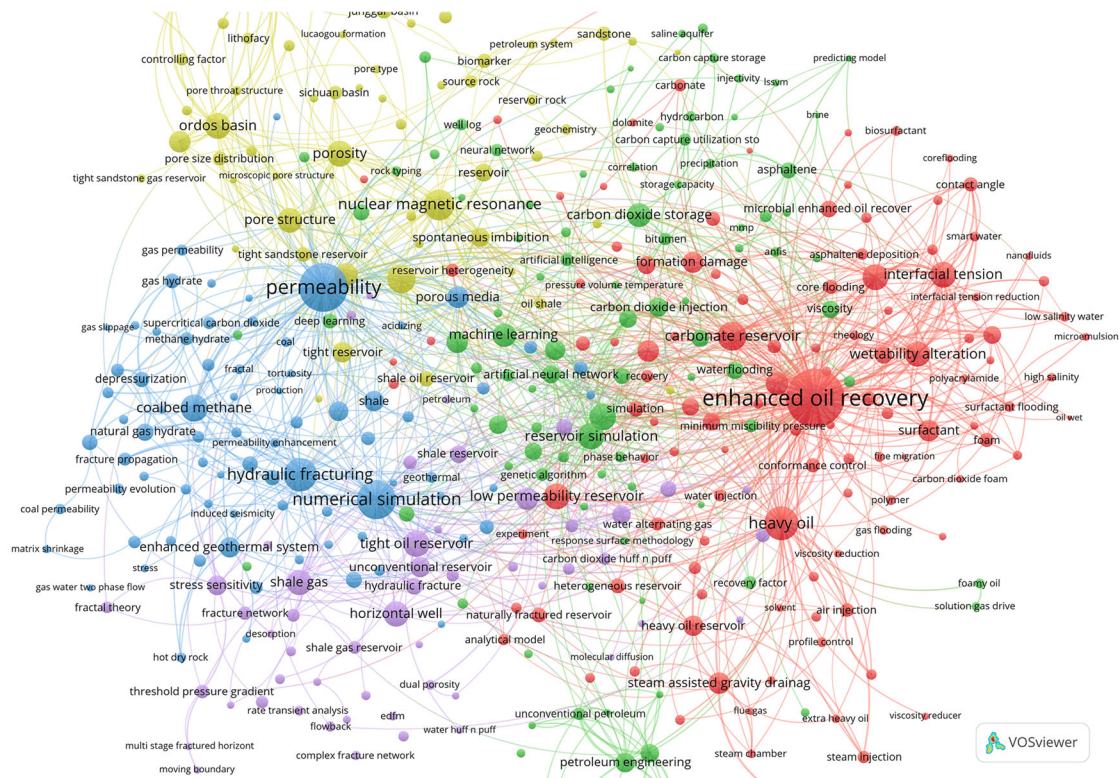


Figure 2. Fragment of the co-occurrence network of normalized author keywords.

There are notable similarities between Figs. 1 and 2. For example, the most commonly used terms “enhanced oil recovery”, “permeability” could still serve as the names of their clusters, but the term “heavy oil” got into the red cluster during normalization, while in the first figure it is in the purple cluster.]

For detailed review of this network, the file AuKWGr4Top400Min50norm.json placed in the archive can be used at <https://app.vosviewer.com/>.

Let’s compare terms in closely related clusters of author keywords obtained by the above two methods.

Note. In the table headings we will use AuKWs to abbreviate “Author Keywords”, and N for occurrence of a given author keyword.

Table 1. Comparison of two clusters, conventionally labeled as “permeability”.

Normalized AuKWs	N	Original AuKWs	N
permeability	392	permeability	392
numerical simulation	250	numerical simulation	245
hydraulic fracturing	187	hydraulic fracturing	187
coalbed methane	125	shale gas	104
porous media	80	coalbed methane	99
enhanced geothermal system	71	horizontal well	88
shale	69	porous media	80
fracture	66	shale	63
natural gas hydrate	51	stress sensitivity	63
two phase flow	44	fracture	48
natural fracture	42	natural gas hydrate	48
tight gas reservoir	42	adsorption	45

depressurization	41	tight oil reservoirs	45
effective stress	40	depressurization	41
gas hydrate	38	effective stress	40
geothermal energy	35	hydraulic fracture	40
temperature	32	two-phase flow	37
mathematical model	31	geothermal energy	35
stimulated reservoir volume	30	enhanced geothermal system	34
finer migration	29	temperature	32
fracture propagation	29	mathematical model	31
gas production	28	apparent permeability	29
methane hydrate	28	finer migration	29
fracturing fluid	27	tight gas reservoir	29
pore network model	27	fracture propagation	28
coalbed methane reservoir	26	gas hydrate	28
modeling	26	gas production	28
supercritical carbon dioxide	26	stimulated reservoir volume	28
finite element method	25	low permeability reservoir	27
permeability anisotropy	24	methane hydrate	26

The terms “permeability”, “numerical simulation”, “hydraulic fracturing” having the same spelling occur in both columns with the same frequency. But already “coalbed methane” occurs 125 times in the left column and 99 times in the right column. Similarly, “two-phase flow” — 44 and “two-phase flow” — 37 and so on.

Table 2. Comparison of two clusters, conventionally labeled as “reservoir simulation”.

Normalized AuKWs	N	Original AuKWs	N
reservoir simulation	120	relative permeability	116
relative permeability	116	reservoir simulation	115
carbon dioxide storage	96	machine learning	90
machine learning	91	co2 storage	85
heterogeneity	79	heterogeneity	79
carbon dioxide sequestration	72	co2 sequestration	66
petroleum engineering	65	petroleum engineering	65
carbon dioxide injection	61	oil/gas reservoirs	58
oil/gas reservoir	58	sensitivity analysis	57
sensitivity analysis	57	capillary pressure	55
artificial neural network	55	viscosity	52
capillary pressure	55	waterflooding	52
carbon dioxide enhanced oil recovery	52	co2 injection	50
viscosity	52	optimization	50
waterflooding	52	asphaltene	44
optimization	50	gas injection	43
asphaltene	45	reservoir characterization	43
in situ combustion	43	artificial neural network	41

reservoir characterization	43	carbon dioxide	40
history matching	40	deep learning	38
deep learning	38	history matching	35
recovery factor	29	co2	32
artificial intelligence	28	asphaltene precipitation	31
genetic algorithm	28	unconventional reservoir	31
minimum miscibility pressure	28	tight reservoirs	28
bitumen	27	bitumen	27
carbon capture utilization storage	27	ccus	27
numerical modeling	27	numerical modeling	27
unconventional petroleum	27	oil recovery factor	27
crude oil	26	unconventional petroleum	27

In the original records, the term “reservoir simulation” was referred to in both singular and plural, which caused the observed difference in the data columns; even more frequently, the term “artificial neural network” appears in both singular and plural.

On the other hand, terms with the single spelling “heterogeneity,” “viscosity,” “deep learning,” “numerical modeling,” and “bitumen” occur the same number of times in both columns.

An example of a significant difference in occurrence is the term “gas injection”, which appears on the right side of Table 2 and on the left side of Table 4.

The term “tight reservoir” does not appear in the cluster of normalized keywords, but it appears in the right-hand column in the plural “tight reservoirs”.

The term “waterflooding” appears 52 times in both columns of this table and in the spelling “water flooding” in Table 3. This term was not normalized in the process of preparing the author's keywords.

Table 3. Comparison of two clusters, conventionally labeled as “enhanced oil recovery”.

Normalized AuKWs	N	Original AuKWs	N
enhanced oil recovery	563	enhanced oil recovery	384
heavy oil	194	eor	147
carbonate reservoir	139	wettability	112
interfacial tension	115	wettability alteration	108
low permeability reservoir	114	carbonate reservoir	99
wettability	112	interfacial tension	92
wettability alteration	109	oil recovery	89
oil recovery	89	formation damage	73
steam assisted gravity drainage	78	surfactant	68
surfactant	78	polymer flooding	52
carbon dioxide	77	low permeability	50
formation damage	73	water flooding	47
heavy oil reservoir	68	carbonate reservoirs	40
carbon dioxide flooding	54	enhanced oil recovery (eor)	39
nanoparticle	54	nanoparticles	36
polymer flooding	52	foam	35
water flooding	50	conformance control	34

imbibition	49	contact angle	34
microbial enhanced oil recovery	47	improved oil recovery	32
simulation	41	sweep efficiency	31
naturally fractured reservoir	37	core flooding	30
heterogeneous reservoir	36	low-permeability reservoirs	29
foam	35	recovery factor	29
oil reservoir	35	microbial enhanced oil recovery	25
carbonate rock	34	asphaltene deposition	23
conformance control	34	heavy oil reservoirs	22
contact angle	34	carbonate rock	21
core flooding	34	permeability reduction	21
air injection	32	nanofluid	20
asphaltene precipitation	31	salinity	20

In the right part of the table there are different spellings of the term: “enhanced oil recovery”, “eor”, and “enhanced oil recovery (eor)”, which can significantly affect the clustering results due to the wide use of this term.

The importance of writing the term with and without hyphen can be seen in this table. “low-permeability reservoir” occurs 114 times in the left column and ‘low-permeability reservoirs’ occurs 29 times in the right column, while ‘low-permeability reservoir’ occurs 27 times but already in Table 1.

The term “heavy oil reservoir” occurs 68 times in the left column, and “heavy oil reservoir” 44 times and “heavy oil reservoirs” 22 times in the right column.

The importance of abbreviation to the singular: “nanoparticle” occurs 54 times on the left side, and “nanoparticles” 36 times on the right side.

Table 4. Comparison of two clusters, conventionally labeled as “tight oil reservoir”.

Normalized AuKWs	N	Original AuKWs	N
tight oil reservoir	116	heavy oil	192
horizontal well	107	tight oil reservoir	69
shale gas	107	co2 flooding	52
fractured reservoir	81	imbibition	49
unconventional reservoir	69	heavy oil reservoir	44
stress sensitivity	66	simulation	41
low permeability	58	low-permeability reservoir	40
shale reservoir	52	unconventional reservoirs	38
hydraulic fracture	48	sagd	34
adsorption	45	air injection	32
gas injection	43	geomechanics	32
improved oil recovery	33	fractured reservoirs	30
geomechanic	32	reservoir heterogeneity	30
fracture network	31	threshold pressure gradient	30
threshold pressure gradient	31	multiphase flow	29
shale gas reservoir	30	oil reservoir	28
apparent permeability	29	recovery	26
multiphase flow	29	steam injection	25

carbon dioxide huff n puff	27	water injection	24
diffusion	25	fractal theory	23
pressure transient analysis	25	analytical model	22
fractal theory	23	pressure transient analysis	22
huff n puff	23	heat transfer	20
flowback	19	steam flooding	20
production performance	19	flowback	19
dual porosity	18	numerical model	19
gas condensate	18	production performance	19
non darcy flow	18	thermal recovery	19
flow regime	16	water cut	19
rate transient analysis	16	water saturation	19

The most important thing to note in this table is that the term “heavy oil” is the most frequent term in the right column, but it appears in the left column in Table 3.

The terms “fractured reservoir” and “unconventional reservoir” occur in the singular in the left column, but in the plural in the right column.

The term “shale gas reservoir” does not appear in the right-hand column.

Table 5. Comparison of two clusters, conventionally labeled as “shale oil” and “porosity”.

Normalized AuKWs	N	Original AuKWs	N
nuclear magnetic resonance	156	shale oil	137
shale oil	137	porosity	114
porosity	114	ordos basin	113
ordos basin	113	pore structure	101
pore structure	101	tight oil	92
tight oil	94	nmr	74
tight sandstone	77	tight sandstone	70
tight reservoir	70	nuclear magnetic resonance	69
diagenesis	66	diagenesis	66
spontaneous imbibition	62	spontaneous imbibition	61
reservoir quality	58	reservoir quality	58
reservoir	56	fractured reservoir	51
shale oil reservoir	41	reservoir	50
junggar basin	40	tight reservoir	42
tight sandstone reservoir	37	junggar basin	40
sandstone	34	sandstone	33
biomarker	33	pore size distribution	31
pore size distribution	33	sichuan basin	31
sichuan basin	31	tarim basin	28
tarim basin	28	yanchang formation	28
yanchang formation	28	shale reservoir	27
controlling factor	27	carbonate	26
source rock	26	shale oil reservoir	24
oil shale	23	tight sandstone reservoir	24

lithofacy	22	controlling factors	23
reservoir rock	22	oil shale	22
fractal dimension	21	fractal dimension	21
reservoir characteristic	21	lithofacies	21
sandstone reservoir	21	reservoir characteristics	21
songliao basin	20	reservoir rock	20

This table is interesting because the lack of disclosure of abbreviations (nmr→nuclear magnetic resonance) can cause that the most frequently occurring term is different on the left and right side of the table.

This table lists the names of the basins “Ordos Basin”, “Junggar Basin”, “Sichuan Basin”, “Tarim Basin”, “Songliao Basin” - all these basins are located in China. This is consistent with the large number of Chinese publications in the studied sample of bibliometric records.

Note: Keep in mind that the above tables only summarize the 30 most frequent terms from the full keyword occurrence tables, so the term occurrence amounts may not match.

The comparison of the terms in the tables above helped to illustrate the importance of term normalization on the results of term clustering.

The importance of term normalization will become even more evident in the next section.

Using the Scimago Graphica program to identify promising research tasks

Within the framework of this article, the promising research tasks are assumed to be described by the author’s keywords that occur more often in new publications that have a high citation rate and connection with other terms. Technically, this can be visualized as a slice of the network of co-occurrence of terms represented in the coordinates “Average publication year of the documents in which a keyword occurs” and “Average normalized number of citations”.

Let’s explain the concepts used by quoting from the VOSviewer manual²: “Avg. norm. citations. The average normalized number of citations received by the documents in which a keyword or a term occurs.” “The normalized number of citations of a document equals the number of citations of the document divided by the average number of citations of all documents published in the same year and included in the data that is provided to VOSviewer.”

Note: For a better understanding of the content of the publications selected as examples, their titles and my abbreviated version of the abstract are given in the body of the paper; they are enclosed in quotation marks to emphasize that this is not plagiarism but citation.

Figure 3 shows the graph of the term co-occurrence network constructed for author keywords taking into account their normalization. The following constraints were used to construct the graph in Scimago Graphica software: clusters 1,2,3 out of five obtained for normalized author keywords by VOSviewer software; total_link_strength >= 44; avg_norm_citations >= 1. Degree >= 25. Degree was calculated for the pre-built network, then the data from this network was exported, the co-occurrence network was re-built using the exported data, and then the degree >= 25 filter was implemented to limit the number of terms displayed in the graph. This filter leaves only terms that are well related to other terms in the graph, which is important for identifying a local topic described by a number of co-occurring terms. A analogous result can be obtained by applying the “link strength” and “total link strength” filters to the original graph, but all these parameters are calculated for the network in question, so the resulting values may differ. In this case, the capabilities of Scimago Graphica were demonstrated. Filters are used to display only the terms most relevant to the analysis being performed to improve the readability of the graph in the publication.

Definition: The degree of a node is the number of links it has with other nodes in the network.

² van Eck N.J., Waltman L. Manual for VOSviewer version 1.6.20. URL: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.20.pdf (accessed 20.02.2025).

According to the VOSviewer manual – “Total link strength attribute indicate the total strength of the links of an item with other items.”

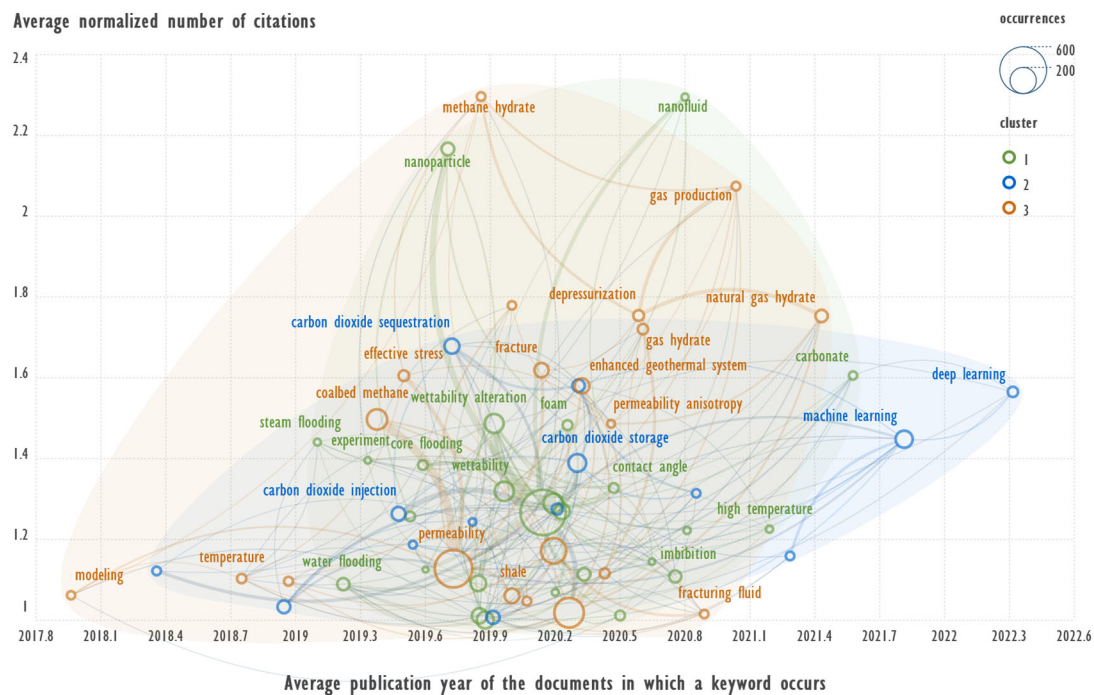


Figure 3. Co-occurrence network of the normalized Author keyword with respect to the above constraints.

Let us consider examples of identifying publications revealing the topics of two clusters (3 and 1) with terms having a high value of the average normalized number of citations.

Third cluster. The results of a sequential search (search in found) for terms appearing in the text fields of all records: methane hydrate → 49 records AND gas production → 33 records AND depressurization → 20 AND fracture → 4. Of the four papers, the most cited is “Enhancement of gas production from methane hydrate reservoirs by the combination of hydraulic fracturing and depressurization method” [15]. According to ScienceDirect, this publication has been cited 171 times as of February 12, 2025. Short summary of annotation: “A fracturing and depressurisation method is proposed to improve the efficiency of gas production from methane hydrate (MG) reservoirs. A model of a fractured MG reservoir was created and the gas production behaviour under different temperature conditions was studied. The effect of increasing fracture zone size and permeability on gas production rate was more prominent in the early stage of depressurisation for the high-temperature reservoir, while the increase in overall gas production was minimal in the low-temperature reservoirs.”

First cluster. The results of a sequential search (search in found) for terms appearing in the text fields of all records: nanofluid → 120 nanoparticle → 93 carbonate → 33 → enhanced oil recovery → 24 → wettability alteration → 16. The most cited paper “Comparative study of using nanoparticles for enhanced oil recovery: Wettability alteration of carbonate rocks” [16], this publication has been cited 261 times as of February 12, 2025. Short summary of annotation: “Exposure of various nanofluids from zirconium dioxide (ZrO2), calcium carbonate (CaCO3), titanium dioxide (TiO2), silicon dioxide (SiO2), magnesium oxide (MgO), aluminium oxide (Al2O3), cerium oxide (CeO2) and carbon nanotubes (CNT) on the wettability of carbonate rocks were investigated for enhanced oil recovery from oil reservoirs. The results of spontaneous imbibition tests and core flooding experiments confirm the active role of CaCO3 and SiO2 nanoparticles in enhanced oil recovery. It is

shown that both irreducible water saturation and inlet capillary pressure increased after treatment with CaCO₃ nanofluid.”

Fig. 4 shows a graph constructed under the same conditions as the previous one, but for the original author's keywords.

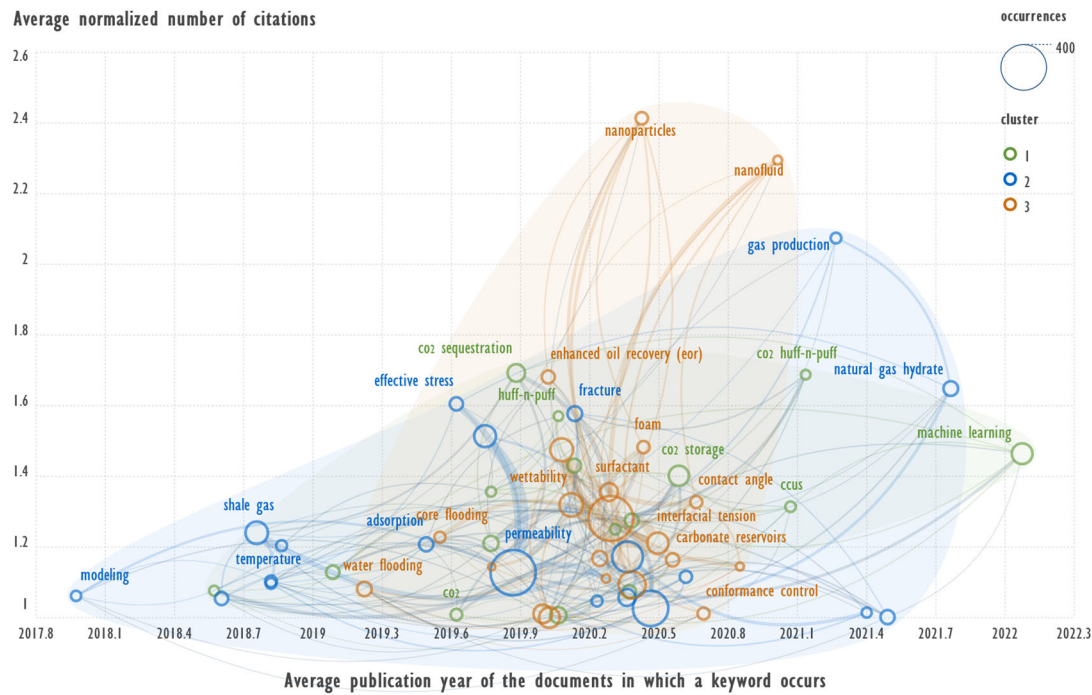


Figure 4. Co-occurrence network of the original Author keyword with respect to the above constraints.

The lack of normalization of term spelling affected the clustering. This is particularly noticeable for the second cluster in this figure. The term *methane hydrate* is not reflected on it, and it is found in articles with high citations according to the previous figure. In the third cluster, the terms *nanoparticles* and *nanofluids* are present, but the term *carbonates*, as in the previous figure, is not present, and it is found in more recent publications. The first cluster in this figure does not contain the term *deep learning*, which is found in newer publications with good citations. If we refer to the interactive web page provided in the archive, we can see those two spellings “enhanced oil recovery” and “enhanced oil recovery (eor)” are presented in the graph. This example illustrates well the importance of pre-processing of text fields on the results obtained.

Co-occurrence network analysis of terms in the text of annotations

The VOSviewer program allows to build a co-occurrence network not only for author and index keywords, but also for terms present in annotation texts. VOSviewer generates a list of noun phrases and brings the terms to the singular number, more details can be found in the manual of the program.

This approach to defining key terms works well, but in the text of abstracts of scientific papers it is quite often found abbreviations of the most commonly used terms in the given research area, e.g. EOR → Enhanced Oil Recovery.

Therefore, the previously compiled list of abbreviations (thesaurus_terms.txt file) was used to construct a network of co-occurrence of terms in the annotation texts. It should be noted that VOSviewer performs whole term replacement, i.e. if in file thesaurus_terms.txt we specified that “eor → enhanced oil recovery”, the term “co2 eor process” will not be replaced, but the individual term “eor” will be replaced by “enhanced oil recovery”. Going beyond this example, it is possible to perform an additional iteration of term normalization by analyzing the terms used to build the

network and expanding the list of replacements in thesaurus_terms.txt. Another way is to perform term substitution in the original annotation text (or titles).

A private observation of the author of this paper shows that the use of textual fields of titles and annotations is underestimated in many publications, the construction of a network of keywords is significantly more frequent. At the same time, the use of title and annotation fields to build a network of terms co-occurrence, in my opinion, gives more coherent clusters and their number is smaller.

The point is that subject matter experts assess the need for further study of an article not by its keywords, but by its title and abstract. The abstract and title fields in the dataset we studied were filled in completely, while the author's keywords had 1470 blank fields out of 8051 total fields.

Another example, exporting from OnePetro downloads titles and abstracts, but not keywords.

The situation is even worse for The Lens database,

on the query — Filters: Year Published = (2024 - 2024) Subject = (Geochemistry and Petrology) received Scholarly Works (12,482). In 12482 records a little more than half (6851) of the annotations are missing and almost all Keywords fields are empty (11911).

Another example is using the “Publish or Perish” program, which can query a large number of resources, including OpenAlex, and the saved files have Title and Abstract fields, but no keywords.

Also, RSS feeds from topical sites have header and body fields, but not keyword fields.

For example, on the old Springer site (link.springer.com/search), you can use RSS with the fields: title and description (which includes an abstract), but not keywords.

According to the text fields of the given examples, it is possible to build a network of term co-occurrence and evaluate the topicality of the collected materials using VOSviewer.

In accordance with the above, Fig. 5 presents the network of terms co-currencies contained in the texts of the abstracts, taking into account the disclosure of abbreviations.

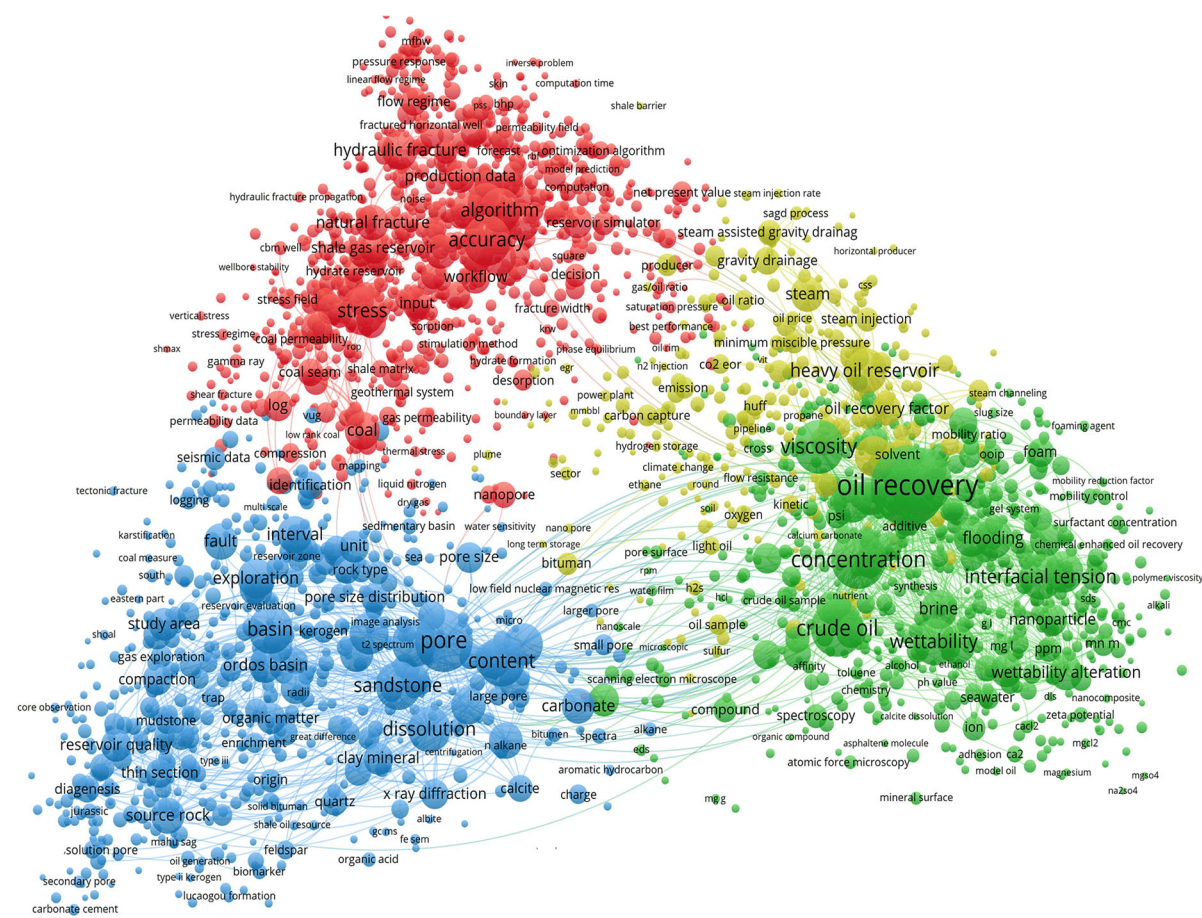


Figure 5. Co-occurrence network of key terms used in the annotation texts, taking into account the disclosure of abbreviations.

First Cluster (Red)

Most frequently appearing terms (Occurrences score): accuracy (499), stress (461), algorithm (384), horizontal well (326), hydraulic fracture (292), reservoir model (288), coal (280).

Terms most frequently used in new publications (Avg. pub. year score): marine hydrate reservoir (2023.1), extreme gradient (2022.6429), seepage field (2022.6), test set (2022.5455), extreme gradient boosting (2022.5), long short term memory (2022.4375), hbs (2022.4)

Here: hbs → hydrate-bearing sediment, the term HBS in annotations this term for the first time can meet with the decoding, and further used only as an abbreviation, it was not included in the list of exceptions, made from the author's keywords.

Terms most frequently used in highly cited publications (Avg. norm. citations score): rock strength (3.1761), slip length (2.8046), co2 ecbm process (2.5662), gas property (2.5647), ugs (2.4861), hydrate bearing layer (2.4357), water permeability (2.4206)

Here: ugs → underground gas storage (UGS), in annotations this term for the first time can meet with the decoding, and further used only as an abbreviation, it was not included in the list of exceptions, made from the author's keywords.

Here: ecbm → enhanced coalbed methane, this term is present in the dictionary of abbreviations compiled using the author's keywords, but it is part of a compound term, so it has not been modified. This feature should be taken into account when compiling the thesaurus_terms.txt file when performing term clustering based on title and abstract texts.

The abbreviations in the following clusters are similar and are not explained further.

Second cluster (green).

Most frequently appearing terms (Occurrences score): oil recovery (1269), enhanced oil recovery (878), concentration (670), crude oil (630), viscosity (560), wettability (422), interfacial tension (410).

Terms most frequently used in new publications (Avg. pub. year score): dispersion stability (2022.6667), promising alternative (2022.4), microfluidic experiment (2022.25), imbibition efficiency (2022.2308), oilwater interfacial tension (2022.1818), md simulation (2022.0833), experimental finding (2022). Here: md → molecular dynamics (MD).

Terms most frequently used in highly cited publications (Avg. norm. citations score): potential solution (2.9926), disjoining pressure (2.8678), viscoelastic property (2.5119), pam (2.4066), nanotechnology (2.1721), early breakthrough (2.1437), co2 foam (2.1364). Here: pam → polyacrylamide (PAM).

Third cluster (blue).

Most frequently appearing terms (Occurrences score): pore (749), content (487), sandstone (460), basin (450), hydrocarbon (434), dissolution (410), pore structure (348).

Terms most frequently used in new publications (Avg. pub. year score): development efficiency (2022.7857), shale oil resource (2022.7727), qingshankou formation (2022.3333), seepage capacity (2022.3214), basalt (2022.2308), movable oil (2022.1579), baikouquan formation (2022.125).

Terms most frequently used in highly cited publications (Avg. norm. citations score): macro pore (2.324), winland (2.2597), pore geometry (2.198), pore distribution (2.185), burial history (2.1773), ct scanning (2.1137), host rock (2.1036). Here: ct → computed tomography (CT).

Fourth cluster (khaki).

Most frequently appearing terms (Occurrences score): heavy oil reservoir (394), heavy oil (326), steam (246), oil recovery factor (197), oil viscosity (188), co2 flooding (158), solvent (154).

Terms most frequently used in new publications (Avg. pub. year score): underground hydrogen storage (2023.2609), hydrogen production (2023.1333), formation energy (2022.6429), nuclear magnetic resonance technology (2022.6), recovery degree (2022.5882), co2 storage efficiency (2022.4615), cushion gas (2022.4545).

Terms most frequently used in highly cited publications (Avg. norm. citations score): cushion gas (3.6855), underground hydrogen storage (3.5665), hydrogen storage (3.5626), hydrogen production (2.897), good potential (2.7012), co2 geological storage (2.317), residual trapping (2.2258).

Explanation: the term “underground hydrogen storage” is more common in new highly cited publications, but the publications themselves are still few, so this term is not reflected in Figs. 6 and 7, where the term occurrence restriction is used.

Application of the Scimago Graphica Program To Identify Promising Research Tasks Using Different Restrictive Filters

Below are the graphs plotted with Scimago Graphica software in the coordinates “Average normalized number of citations” and “Average publication year of the documents in which a keyword occurs” using different restriction filters.

Fig. 6 is plotted using the following sampling constraints: cluster $\rightarrow 1,2,3$; total_link_strength $\rightarrow 100$; occurrences $\rightarrow 100$; avg.*norm.citations* $\rightarrow 1$.

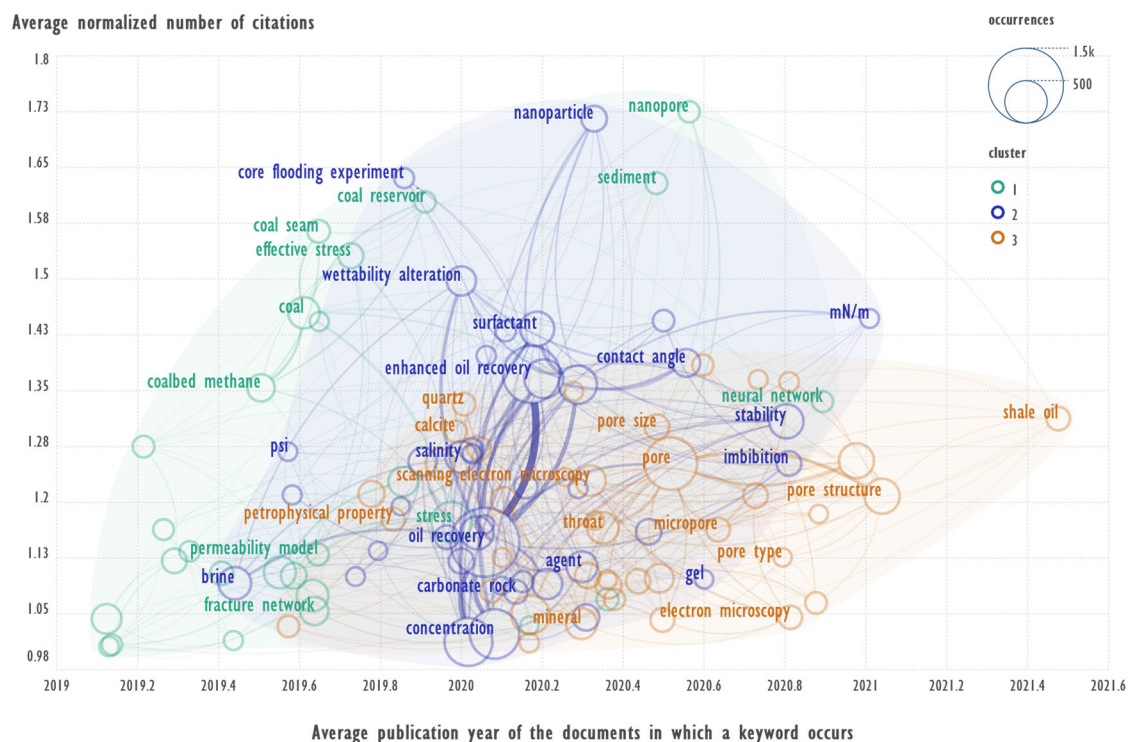


Figure 6. Co-occurrence network of key terms used in the annotation texts, with respect to the above constraints.

The term of the first cluster “nanopore” has the highest number of links with the terms “shale oil” and “pore size” of the third cluster and the term “molecula” of the second cluster. A sequential search for the occurrence of terms in the text fields of all records yields: nanopore $\rightarrow$ 178, shale oil $\rightarrow$ 49, pore size $\rightarrow$ 14, molecula $\rightarrow$ 4. Of the four articles, the most relevant is “Molecular dynamics simulations of oil transport through inorganic nanopores in shale” [17] which has been cited 390 times (14-Feb-2025). Short summary of annotation: “The transport of liquid hydrocarbon through nanopores of inorganic minerals is crucial for the development of fluid-rich shale reservoirs and for understanding oil migration from deep-seated source rocks with extremely low permeability. The authors used non-equilibrium molecular dynamics to study the flow of octane in quartz slits under pressure based on the Navier-Stokes equation with slip boundary and viscosity corrections”.

Note: Leiden algorithm attempts to optimize modularity in extracting communities from networks [18]. “The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random” $\rightarrow$ [19]. Based on the above, the current example is interesting because well-connected terms can belong to different clusters, since the membership of a cluster is determined by the set of links in it.

The fact that the terms “nanopore” and “nanoparticle” belong to different clusters is also interesting; they are close in spelling but have different contexts.

The term “nanoparticle” often occurs with the terms “enhanced oil recovery”, “interfacial tension”, “wettability”, referring already to the same cluster.

A sequential search for the occurrence of terms in the text fields of all records yields: nanoparticle → 248, enhanced oil recovery → 145, interfacial tension → 69, wettability → 53, rock surface → 9.

The fact that a sequential search for the five terms yielded 9 results indicates the stability of the topic they are describing.

Of the nine articles, the most relevant is “Adsorption analysis of natural anionic surfactant for enhanced oil recovery: The role of mineralogy, salinity, alkalinity and nanoparticles” [20] which has 228 citations.

Short summary of annotation: “Anionic surfactants are widely used as effective chemical reagents for enhanced oil recovery. This study deals with the equilibrium adsorption and kinetics of anionic surfactant synthesised from soapnut fruit on sandstone, carbonate and bentonite clay, which are reservoir rocks. The presence of alkali and nanoparticles reduces the loss of surfactant during adsorption and has a synergistic effect in reducing the interfacial tension, which is favourable for the application of surfactant in oil recovery.”

Fig. 7 is plotted using the following sampling constraints: cluster → 1 and 3; total_link_strength → 100; occurrences → 20; avg.norm.citations → 1.5.

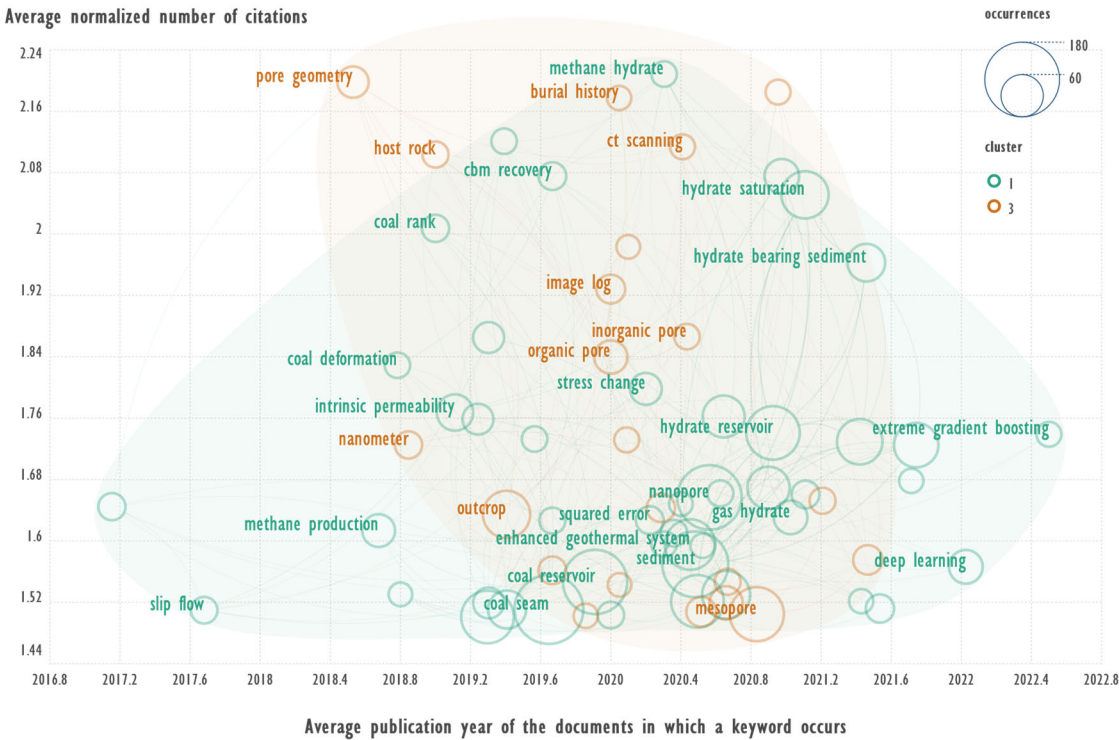


Figure 7. Co-occurrence network of key terms used in the annotation texts, with respect to the above constraints.

This graph shows only the two clusters shown in Fig. 5 — 1 and 3. A significant difference with Fig. 6 is that this data slice includes more rarely occurring terms (occurrences → 20 vs occurrences → 100) and the threshold for the average normalized citation is raised (avg.norm.citations → 1.5 vs avg.norm.citations → 1). Also note the slight shift of the right-most value on the average publication time axis (2022.8 vs 2921.6).

Thus, this slice can be seen as a reflection of the more promising topics: there are not many publications yet, but they are more cited and published more recently.

For the sake of brevity, let us consider only one new topic that appeared in the first cluster and is described by the terms: methane hydrate, hydrate saturation, hydrate bearing sediment.

A sequential search for the occurrence of terms in the text fields of all records yields: methane hydrate → 49, hydrate saturation → 24, hydrate bearing sediment → 5. Note that the number of publications is lower than in the examples in the previous figure.

Publication “Experimental study on the gas phase permeability of methane hydrate-bearing clayey sediments” [21] cited 74 times. Short summary of annotation: “The permeability of hydrate-bearing sediments is one of the important parameters affecting the rate of gas production in hydrate reservoirs. In this paper, a series of experiments were conducted to investigate the gas-phase permeability of kaolin clay under different hydrate saturation. The results showed that the gas-phase permeability of kaolin clay firstly decreases and then increases with increasing hydrate saturation. The gas-phase permeability of hydrate-saturated clay samples at high effective axial stress is lower than that at low stress, which is due to pore space compaction.”

It was interesting to find the area of application of the method “extreme gradient boosting”; such publications turned out to be 18 such publications.

It was interesting to find the area of application of the “extreme gradient boosting” method. 18 such publications were found.

The most interesting article [22] has been cited 86 times. However, none of the 18 publication abstracts contained the term “hydrate”, suggesting that the inclusion of the term “extreme gradient boosting” in this cluster is due to co-occurrence with other terms, such as “permeability”, but not “hydrate”. The term “permeability” may be relevant in the context of records relating to “cbm recovery” which in turn has a relationship to the term “methane hydrate”.

This example shows that a terms co-occurrence network, although a clear method for identifying relevant or promising research tasks, is not always convenient for extending literature collection by composing queries from co-occurring terms. For query generation, the direct method of determining the co-occurrence of terms (by algorithms like Apriori) may be more promising than assigning terms to a single cluster.

Conclusions

Despite the fact that the Index keyword used in a query to Scopus is rarely found in author keywords and abstracts, bibliometric records exported by such a query are relevant to the given topic.

Examples are given to show the importance of pre-processing text fields before performing term clustering.

The expediency of disclosing the abbreviations of key terms when constructing the terms co-occurrence network encountered in the abstracts of publications by the VOSviewer program is shown.

The effectiveness of using filters in the Scimago Graphica program to construct a term co-occurrence network to identify promising research topics is demonstrated.

It has been observed that in some cases terms occurring in the same cluster are not the best choice for querying to expand the collection of publications on a given topic. It is proposed to conduct a separate study on the use of Apriori class algorithms for this purpose.

Promising research tasks can be described by the following terms according to the bibliometric data used: 1. nanopore, shale oil, pore size, molecular; 2. nanoparticle, enhanced oil recovery, interfacial tension, wettability, rock surface; 3. methane hydrate, hydrate saturation, hydrate bearing sediment.

Funding: the work was funded by the Ministry of Science and Higher Education of the Russian Federation No. 122022800270-0.

Note to the editor: To make it clear to the reader why a particular article has been selected as relevant or promising based on a number of key terms, the titles of the selected publications and a shortened version of their abstracts are given in the text enclosed in quotation marks. Anti-plagiarism programs will probably respond to

this, but without showing at least some of the text, the reader may not understand why the article was selected. (delete when publishing).

References

1. Guthrie S., Ghiga I., Wooding S. What do we know about grant peer review in the health sciences? // *F1000Res*. 2018. Vol. 6. P. 1335. <https://doi.org/10.12688/f1000research.11917.2>.
2. Recio-Saucedo A., Crane K., Meadmore K., et al. What works for peer review and decision-making in research funding: a realist synthesis // *Res Integr Peer Rev*. 2022. Vol. 7, No 1. P. 2. <https://doi.org/10.1186/s41073-022-00120-2>.
3. Moran R., Butt J., Heller S., et al. Health research systems in change: the case of 'Push the Pace' in the National Institute for Health Research // *Health Res Policy Sys*. 2019. Vol. 17, No 1. P. 37. <https://doi.org/10.1186/s12961-019-0433-2>.
4. Van Eck N.J., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // *Scientometrics*. 2010. Vol. 84, No 2. P. 523–538. <https://doi.org/10.1007/s11192-009-0146-3>.
5. Ravi Kumar N., Kulkarni P., Kalaiarasai V. Unveiling the Viral Thread: A Comprehensive Analysis of Virality Coefficients in Indian Fashion Brand Dynamics // *Advances in Marketing, Customer Relationship Management, and E-Services* / ed. Tarnanidis T.K., Papachristou E., Karypidis M., et al. IGI Global, 2024. P. 180–193. <https://doi.org/10.4018/979-8-3693-3049-4.ch011>.
6. Van Veldhoven Z., Etikala V., Goossens A., et al. A Scoping Review of the Digital Transformation Literature Using Scientometric Analysis // *Bus. Inf. Sys*. 2021. P. 281–292. <https://doi.org/10.52825/bis.v1i.49>.
7. Chigarev B. A Proof-of-Concept Methodology for Identifying Topical Scientific Issues in New Publications Whose Citations Have Not Yet Been Established. 2024. <https://doi.org/10.20944/preprints202410.2220.v1>.
8. Chigarev B. Analyzing the Possibilities of Using the Scilit Platform to Identify Current Energy Efficiency and Conservation Issues. 2024. <https://doi.org/10.20944/preprints202404.0744.v1>.
9. Handayani M., Sunarya M.H., Bahit M. Visualization of Fintech Research Trends (Financial Technology) Using VOSViewer // *Proceedings of the 3rd Annual Management, Business and Economics Conference (AMBEC 2021)* / ed. Eltivia N., Riwajanti N.I., Susilowati K.D.S. Dordrecht: Atlantis Press International BV, 2023. Vol. 224. P. 3–10. https://doi.org/10.2991/978-94-6463-026-8_2.
10. Chigarev B. GSDMM Clustering Results Visualization Technique for Short Texts. 2024. <https://doi.org/10.20944/preprints202412.0705.v1>.
11. Feldman R., Sanger J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. 1st ed. Cambridge University Press, 2006. <https://doi.org/10.1017/CBO9780511546914>.
12. Cammarano A., Varriale V., Michelino F., et al. Discovering technological opportunities of cutting-edge technologies: A methodology based on literature analysis and artificial neural network // *Technological Forecasting and Social Change*. 2024. Vol. 209. P. 123811. <https://doi.org/10.1016/j.techfore.2024.123811>.
13. Gitelman L., Magaril E., Kozhevnikov M. Energy Security: New Threats and Solutions // *Energies*. 2023. Vol. 16, No 6. P. 2869. <https://doi.org/10.3390/en16062869>.
14. Hassan-Montero Y., De-Moya-Anegón F., Guerrero-Bote V.P. SCImago Graphica: a new tool for exploring and visually communicating data // *EPI*. 2022. P. e310502. <https://doi.org/10.3145/epi.2022.sep.02>.
15. Feng Y., Chen L., Suzuki A., et al. Enhancement of gas production from methane hydrate reservoirs by the combination of hydraulic fracturing and depressurization method // *Energy Conversion and Management*. 2019. Vol. 184. P. 194–204. <https://doi.org/10.1016/j.enconman.2019.01.050>.
16. Nazari Moghaddam R., Bahramian A., Fakhrouieian Z., et al. Comparative Study of Using Nanoparticles for Enhanced Oil Recovery: Wettability Alteration of Carbonate Rocks // *Energy Fuels*. 2015. Vol. 29, No 4. P. 2111–2119. <https://doi.org/10.1021/ef5024719>.
17. Wang S., Javadpour F., Feng Q. Molecular dynamics simulations of oil transport through inorganic nanopores in shale // *Fuel*. 2016. Vol. 171. P. 74–86. <https://doi.org/10.1016/j.fuel.2015.12.071>.
18. Waltman L., Van Eck N.J. A smart local moving algorithm for large-scale modularity-based community detection // *Eur. Phys. J. B*. 2013. Vol. 86, No 11. P. 471. <https://doi.org/10.1140/epjb/e2013-40829-0>.

19. Newman M.E.J. Modularity and community structure in networks // Proc. Natl. Acad. Sci. U.S.A. 2006. Vol. 103, No 23. P. 8577–8582. <https://doi.org/10.1073/pnas.0601602103>.
20. Saxena N., Kumar A., Mandal A. Adsorption analysis of natural anionic surfactant for enhanced oil recovery: The role of mineralogy, salinity, alkalinity and nanoparticles // Journal of Petroleum Science and Engineering. 2019. Vol. 173. P. 1264–1283. <https://doi.org/10.1016/j.petrol.2018.11.002>.
21. Liu W., Wu Z., Li Y., et al. Experimental study on the gas phase permeability of methane hydrate-bearing clayey sediments // Journal of Natural Gas Science and Engineering. 2016. Vol. 36. P. 378–384. <https://doi.org/10.1016/j.jngse.2016.10.055>.
22. Zhong R., Johnson R., Chen Z. Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost) // International Journal of Coal Geology. 2020. Vol. 220. P. 103416. <https://doi.org/10.1016/j.coal.2020.103416>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.