Article

A Proof-of-Concept Methodology for Identifying Topical Scientific Issues in New Publications Whose Citations Have Not Yet Been Established

Boris Chigarev *

Posted Date: 29 October 2024

doi: 10.20944/preprints202410.2220.v1

Keywords: topical research issues; keyword weight estimation; lemmatization; VOSviewer; Scimago Graphica; industry digitalization

Article

A Proof-of-Concept Methodology for Identifying Topical Scientific Issues in New Publications Whose Citations Have Not Yet Been Established

Boris Chigarev

Oil and Gas Research Institute of the Russian Academy of Sciences (OGRI RAS), Moscow, Russia; bchigarev@ipng.ru

Abstract: Identification of topical research issues using bibliometric data is complicated by the fact that the citation of publications from recent years has not yet been formed. In this paper, it is proposed to use the average citation of the journal over two years rather than the article citation to estimate to estimate the weight of the keyword occurring in the sample under consideration. In order to identify the terms that characterize relevant research topics, it is proposed to represent the term co-occurrence network in coordinates of the average occurrence of the term per year and the average normalized citation of the term to visualize the graph. Furthermore, this methodology proposes the use of preprocessing of keywords using a lemmatization dictionary. 3,696 bibliometric records for 2022–2024 from the ScienceDirect platform on the topic of industry digitalization were used for the analysis. The VOSviewer and Scimago Graphica programs were used sequentially. The former was used to display the overall landscape of the study, while the latter was used to analyze in more detail the individual slices of bibliometric data obtained with VOSviewer. A 'convex hull' was used to facilitate the perception of cluster boundaries. After analysing the data and highlighting the terms, it is proposed to provide context by quoting strings from publications and defining of lesser-known terms. The industry digitalization is not only a technical and technological issue but also an economic one, as evidenced by terms such as 'digital economy' and 'Industry 5.0'.

Keywords: topical research issues; keyword weight estimation; lemmatization; VOSviewer; Scimago Graphica; industry digitalization

Introduction

Motivation and Relevance of This Study

1. In determining the actual tasks of research using bibliometric data, one is inevitably confronted with the fact that the citability of publications in recent years has not yet been established. Therefore, the 'Cited by' field or other similar fields of bibliometric data exported from abstract databases are still poorly populated. For example, by query: 'q = (title:(digital energy) OR abstract:(digital energy) OR keyword:(digital energy) OR field_of_study:(digital energy)) &p=0&n=10&publicationType.must=journal article&publicationType.must=conference proceedings article&publishedDate.from=2022–10–01&publishedDate.to=2024–10–01' to the abstract database The Lens of the 9223 bibliometric records obtained, 5570 publications have not yet been cited and 1359 have been cited only once. Therefore, using the indicator such as the average normalized citation utilized by the most frequently used in bibliometric analysis program VOSviewer [1] will not be relevant (in the sense of correspondence of the obtained result to the desired result). Based on the above, this methodology proposes to use the average citation of a journal over two years (Cites / Doc. (2years) as by Scimago Journal & Country Rank) instead of article citations to estimate the weight of a keyword occurring in the sample under consideration. This study used bibliometric data from the ScienceDirect platform and the file 'scimagojr 2023.csv' containing the characteristics of journals published by Elsevier.

doi:10.20944/preprints202410.2220.v1

2. The VOSviewer program provides high–quality results of the bibliometric analysis, including a visual representation of the average occurrence of a term per year and the average normalized citation of a term. However, in my opinion, in order to identify terms characterizing topical research themes, it is desirable to present a diagram of occurrence of relevant terms simultaneously in both the above–mentioned coordinates. Various methods of Layout: Force Directed, Kamada and Kawai's algorithm, LinLog, Dagre Top–Down have been used to visualize the co-occurrence network of terms, but I have not been able to find publications by other authors that have jointly used the coordinates of the average term occurrence per year and the average normalized citation of the term to visualize the graph layout.

3. The VOSviewer program allows using thesaurus_terms.txt file to replace the spelling of terms, but unfortunately this feature is rarely used. In this paper, we propose to use this feature to replace the abbreviations found in keywords with their full spelling, e.g. IoT → internet of things. Author keywords also occur in different spellings, e.g. a term can be used in both plural and singular. Therefore, this technique proposes the use of pre-lemmatization of keywords using a lemmatizer dictionary. This type of lemmatizer is chosen as the most transparent in terms of the changes it makes and can be easily supplemented with new entries, e.g. some complex terms can have both a merged spelling and a separated or hyphenated spelling.

Notes

The VOSviewer program was chosen as the base program, because it directly provides the features necessary to implement the proposed methodology. It is widely used in bibliometric analysis and is, in my opinion, the most consistent (by mutual consistency) and widely used. An example of Scilit database queries reflecting the use of free programs in bibliometric research: 'VOSviewer' for common fields [Title, Abstract, Keyword] gives 11530 publications, Bibliometrix – 2373, CiteSpace – 9860 in publications for 2022–2024. CiteSpace and Bibliometrix provide very diverse capabilities of different methods, so a separate study is needed on how they can be used to implement the methodology described in this paper, for example, to generate a different measure of keyword weights from the available data for the journals in whose publications they occur.

Despite the fact that the lemmatization used a dictionary of 1060 terms related to the considered topic (industry digitalization) and the application of abbreviations transcripts, the results of the analysis showed the incompleteness of the necessary substitutions. This fact is used to demonstrate the importance of normalizing Author keywords before conducting a bibliometric analysis to identify promising research topics.

The choice of 'industry digitalization' as a subject area is due to my professional interests.

The article is written in 'proof of concept' format and does not pretend to be a comprehensive review of the issue of finding a replacement for average normalized citations (Avg. norm. citations as by VOSviewer), but considers only one option, in which averaging by publication is replaced by averaging by journal for the two preceding years. This option is available in the Scimago Journal & Country Rank data. Alternatively, one could, for example, consider the Hirsch index for the journal in which the term occurs in the generated sample for the query of interest, restricted to recent years.

This preprint is intended to be posted on the preprints.org platform[1], which allows comments. Therefore, I would be grateful for comments with links that at least partially discuss a methodology similar to the one proposed in this paper. So far, I have not been able to find an article that fully addresses all of the above issues, but only selected ones. The references found are intended to be used in writing a journal article.

Brief Literature Review

Quality assessment of publications is a topical issue and is addressed in many papers, whose authors may have significantly different opinions.

[1] https://www.preprints.org — The Multidisciplinary Preprint Platform

In [2] it is noted that determining the importance of an article is a complex and time–consuming task. Formal citations are considered the 'gold standard' in the scientific literature, but citation metrics have their own problems. The main problem is the time lag in citations, which makes them insufficient for evaluating recently published papers.

Thus, the authors of [3] explore three methods for evaluating the quality of a scientific article: subjective post–publication peer review, the number of citations, and the impact factor of the journal, with the impact factor being the most satisfactory.

There is another challenge of using formal metrics, as [4] notes that many researchers tend to publish their work in journals with the highest JIF in their field. The Journal Impact Factor (JIF) is a widely used metric that measures a journal's impact by calculate the total number of citations received by the journal in that year for articles published over the previous 2 years divided by the total number of citable items published by the journal in that 2–year period.

The abundance of open access scientific articles online has made it easier for scientists and clinicians to stay updated on research activity [5]. However, this has made it more challenging to find high–quality, relevant articles and journals. To evaluate this, researchers use citation metrics, usage metrics, and alternative metrics (so–called altmetrics).

Thus, even this brief overview shows that there is not only a diversity of evaluations of publications and journals, but also a variety of challenges in applying these evaluations.

In addition, depending on the task, not only the evaluation of articles is important, but also, for example, the evaluation of the relevance of keywords for the generation of queries, e.g. for the generation of systematic reviews.

Author keywords play a crucial role in the scientific literature, influencing indexing terms. Understanding changes in their occurrence over time can provide insight into the evolution of a discipline and be of interest for bibliometric analysis.

One of the significant challenges is the normalization of authors' keywords. The paper [6] proposes a method to extract domain keywords and build an efficient domain thesaurus for better description and analysis of domain news. It improves domain efficiency by combining key information and building a high–quality thesaurus through manual analysis and automated processing.

The proposed methodology for identifying relevant research topics based on clustering of keywords of authors of new publications uses the capabilities of the VOSviewer program, the results of which depend on the compilation of a thesaurus that normalizes the spelling of authors' keywords.

There are few works in which the thesaurus is used when working with VOSviewer; I managed to find only three. At the same time, the thesaurus compilation itself is not formalized in these works, which increases the probability of errors and incomplete filling of the file.

In [7], Figure 4 shows the use of a "Top of the keywords list after data cleaning and thesaurus of terms".

The article [8] also uses the dictionary substitution shown at "Figure 2. (a) A list of merging synonyms, spelling differences, and plurals used during conducting the keyword co-occurrence analysis".

The author of this paper has already used substitute spelling of keywords in the preprint [9].

The above has prompted the compilation of a methodology for analyzing keyword co-occurrence in a proof-of-concept format that addresses these issues.

The Objective of the Paper

To compile a methodology in 'proof of concept' format to identify relevant research topics from bibliometric data of the past two years and the current year on 'industry digitalization' based on open data provided by Elsevier Publishers.

doi:10.20944/preprints202410.2220.v1

Materials and Methods

Bibliometric data for 2022–2024 were exported from the open access platform ScienceDirect for the query: 'Title, abstract, keywords: (industry OR manufacture OR fabrication) AND digitalization', yielding 3,696 results. Current as of 30–09–2024.

Data were exported in RIS format and further translated into CSV format. Field headers were renamed according to their spelling for similar data fields exported from the Scopus database. This procedure is necessary to import the data into the VOSviewer program. Estimation of citations for two years (Cites / Doc. (2years) as by Scimago Journal & Country Rank) was taken from the file 'scimagojr 2023.csv' containing the main characteristics of journals from the SJR platform[2].

This estimate was reduced to a form similar to the citation of publications: the fractional value from the file 'scimagojr 2023.csv' was multiplied by 10 and reduced to integers. That is, in the data in Scopus CSV format, the 'Cited by' field was replaced exactly by the proposed citation score, not by the citation itself.

The data were joined by the journal name field. In addition, data from the 'SJR Best Quartile' field were included, reflecting which quartile the journal belongs to. During the analysis it was used to compare the terms reflecting the current research topics in the journals included in Q1 and others.

There was some inconsistency between the names given in the file 'scimagojr 2023.csv' and the data in the RIS files. The spelling of the mismatched names was checked manually. They were most often reduced to the use of '&' instead of 'and' and the use of an abbreviated name in the full journal name in one case and no abbreviation in another. The fields reflecting the ISSN in both sources were also filled in differently. In this particular case, combining by the journal name field gave the best results and was used.

Some of the journals in the RIS files were not found in the list of journals and their properties file. The main problem was that these were new journals for which data had not yet been generated, and they were not listed in the 2023 journal list. This inconsistency can be used to find new journals in the topic of interest, but that is beyond the scope of this preprint.

The routine work of harmonizing the fields is not reflected in this publication so as not to distract the reader from the main purpose of the work.

The entries of the 'Author Keywords' field were preprocessed. Hyphen in compound words was removed (as there were some keywords without it). Abbreviations in brackets were excluded, abbreviations as separate keywords were replaced by their full name containing 366 substitutions, for example, 3dcp → 3d concrete printing, 3dpc → 3d printed concrete, iomt → internet of medical things, fea → finite element analysis, rve → representative volume element, cdm → continuum damage mechanics, bim → building information modelling, ipd → integrated project delivery, nn → neural network, am → additive manufacturing. This file was composed of abbreviations and full names. The resulting substitutions could exceed the list of separately occurring keywords as abbreviations, but reduced the risk of missing substitutions.

Keywords were subjected to dictionary lemmatization. The dictionary was built from the intersection of a generic lemmatization dictionary consisting of 246278 entries obtained by data collection on github and manually augmented as individual studies were conducted, and a list of unique one–word terms obtained from the Author keywords list. After some manual editing, the dictionary contained 1060 entries.

Even with this preparation of keywords, the results of their analysis retained some differences in spelling. For example, in one case the words of a complex term were connected by a short dash rather than a hyphen.

The thus prepared bibliometric data in Scopus SCV format and abbreviation replacement file were imported into VOSviewer. The parameters used in the clustering are indicated in specific places in the section 'Results and Discussion'.

[2] https://www.scimagojr.com/ — Scimago Journal & Country Rank

This paper presents only the results of clustering the co-occurrence of pre-prepared author keywords. Three analyses were performed: for the full list of records, for the list of records related to Q1 journals, and for the list of the remaining non–Q1 journals.

The clustering results data exported from VOSviewer were used in Scimago Graphica [10] to present them as separate diagrams in the coordinates 'Avg. pub. year' vs 'Avg. norm. citations' for each cluster detected in VOSviewer.

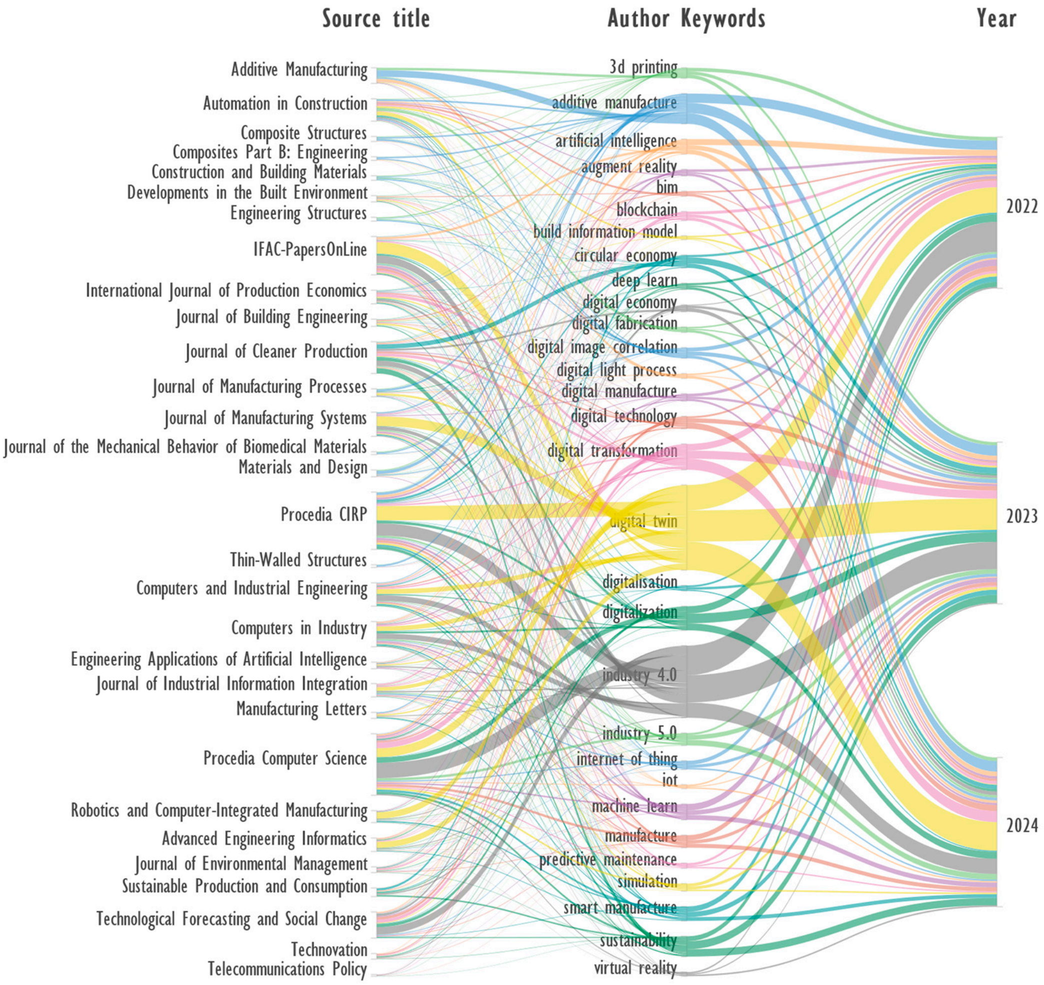
A 'Convex hull' was used to facilitate the perception of the clusters obtained by the algorithm included in Scimago Graphica. The clustering algorithm from Scimago Graphica was configured to obtain 4 clusters. Filtering of the number of terms in the graph was done by limiting the parameters 'total link strength' and 'Avg. norm. citations'.

Adjustment of labels location on the graph was performed by Inkscape [3] program using SVG data exported from Scimago Graphica.

To clarify the significance of terms describing promising research topics, brief summaries of publications corresponding to these terms were provided.

Results and Discussions

To get an overview of the occurrence of Author keywords across journals and across years, we took the 30 journals with the highest number of publications, selected the 30 most frequent terms, and plotted the Sankey diagram shown in Figure 1.



[3] https://inkscape.org/ — Inkscape is a free and open-source vector graphics editor

Figure 1. Distribution of top 30 Author keywords by top 30 journals and years.

A diagram of this type allows to quickly compare selected keywords and the journals in which they are most often found.

The study will further focus on analyzing the co-occurrence of keywords and their assessment based on the average citations of journal articles in the previous 2 years.

Author Keywords Co-Occurrence Analysis for All Journals in the Sample

Figure 2 shows the results of clustering based on co-occurrence of Author keywords obtained by VOSviewer using the following parameters: the occurrence threshold for terms is 3, the number of terms for plotting is 500, the minimum number of terms in a cluster is 60. In this case, 5 clusters were obtained. In total there were 10300 Author keywords, 984 meet and more 3 times.

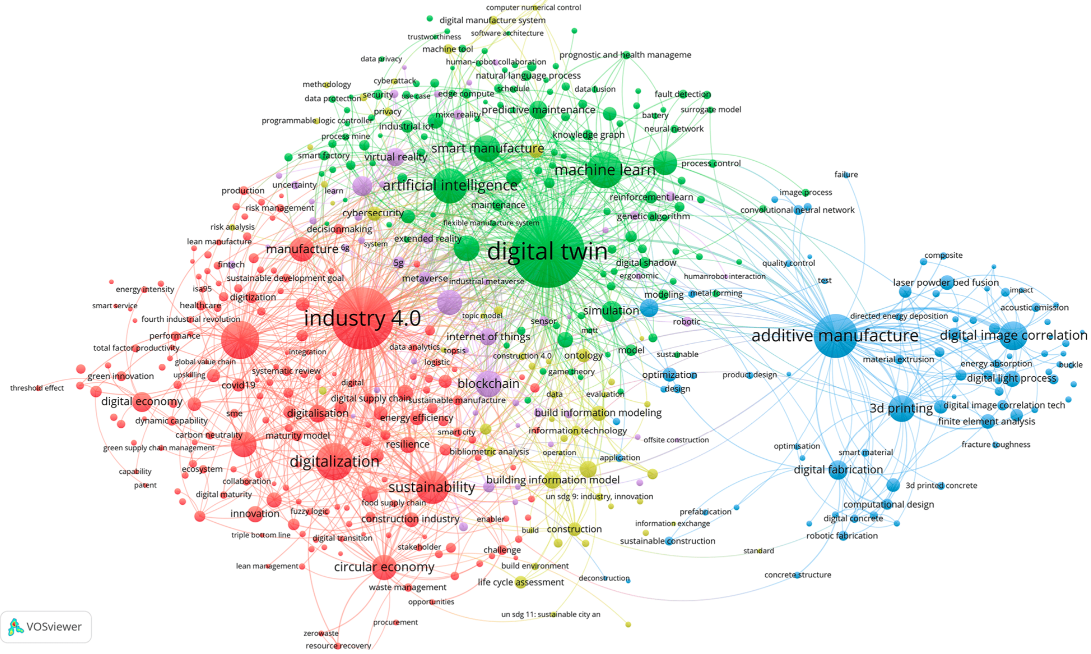


Figure 2. Author keywords co-occurrence network for all journals in the sample.

Table 1 shows 15 Author keywords typical for new publications. It is obtained by sorting the field score<Avg. pub. year>. For ease of reading, the tables have abbreviated headings, e.g., score<Avg. pub. year> → Avg.pub.year.

Table 1. Author keywords typical for new publications.

label	cluster	Occurrences	Avg.pub.year	Avg.norm.citations
incentive mechanism	2	5	2024	1.1297
education	5	7	2023.857	0.4853
modular construction	5	6	2023.833	1.3724
explainable ai	2	5	2023.8	1.1323
industrial metaverse	5	5	2023.8	1.6626
data privacy	2	4	2023.75	0.9745
green supply chain management	1	4	2023.75	1.4205
humancentric	5	4	2023.75	1.6294

operation management	1	4	2023.75	1.689
opportunities	1	4	2023.75	0.7481
strategic management	1	4	2023.75	0.4743
vat photopolymerization	3	11	2023.727	1.486
stakeholder	1	7	2023.714	1.0619
capability	1	3	2023.667	1.5551
cognitive digital twin	2	6	2023.667	1.1499

The term 'incentive mechanism' is typical of the newest publications.

Citation from paper [11] revealing the context for the term 'incentive mechanism': "Under the decentralized decision–making mode, the incentive mechanism positively promotes the innovation efforts of both parties of cooperative enterprises, while the local government is not affected, while under the centralized decision–making mode, the incentive mechanism positively affects the innovation efforts and innovation benefits of all three parties".

Table 2 shows 16 Author keywords typical for publications with the highest average citation. Obtained by sorting the field score<Avg. norm. citations>.

Table 2. Authors keywords of publications with the highest average citation rate.

label	cluster	Occurrences	Avg.pub.year	Avg.norm.citations
battery	2	5	2023.6	2.675
information fusion	2	4	2023.25	2.5233
digital concrete	3	9	2023	2.0635
energy transition	1	9	2023.222	2.0086
electronic	1	4	2022.75	1.9497
ambidexterity	1	3	2022.333	1.9495
patent	1	3	2022.667	1.8555
plan	2	5	2022.6	1.8401
service innovation	1	5	2022.2	1.7643
wire arc additive manufacturing	3	3	2023.333	1.7564
multicriteria decisionmaking	1	6	2022.667	1.7421
software architecture	2	3	2022.667	1.7049
operation management	1	4	2023.75	1.689
value capture	1	6	2023.333	1.6801
resourcebased view	1	4	2023	1.6779
3d concrete printing	3	24	2022.917	1.6702

The terms 'digital concrete' and '3d concrete printing' often appear in highly cited journals, so let's give the context of their use [12]. "Digital concrete technologies, such as 3D concrete printing, have the potential to be game changers in the construction industry by reducing material consumption and offering a high level of automation."

Diagrams Creation with Scimago Graphica Program

Figures 2x1–5 were plotted using the program Scimago Graphica for each of the 5 clusters shown in Figure 2. Technically, this was done by filtering the data by cluster parameter. In order to reduce the number of terms in the graph and to obtain a connected graph, an additional filter was applied to filter out terms that had a 'total link strength' parameter less than or equal to 30. The 'cluster 1–4' in the legend on the graphs reflect the grouping of terms implemented by the algorithm built into Scimago Graphica. They should not be confused with the clusters of Figure 2, as they are an additional grouping of terms. The use of the 'Convex hull' allows to more clearly highlight the terms that are on its boundary, i.e. that take extreme values for a given group of terms.

The term networks in Figures 2.1–5 were constructed without using any layout method, in the coordinates 'Avg.pub.year' vs 'Avg.norm.citations', which facilitated the selection of terms that could be interpreted as reflecting topical research issues. We could not find examples of this kind of term network construction in the works of other authors.

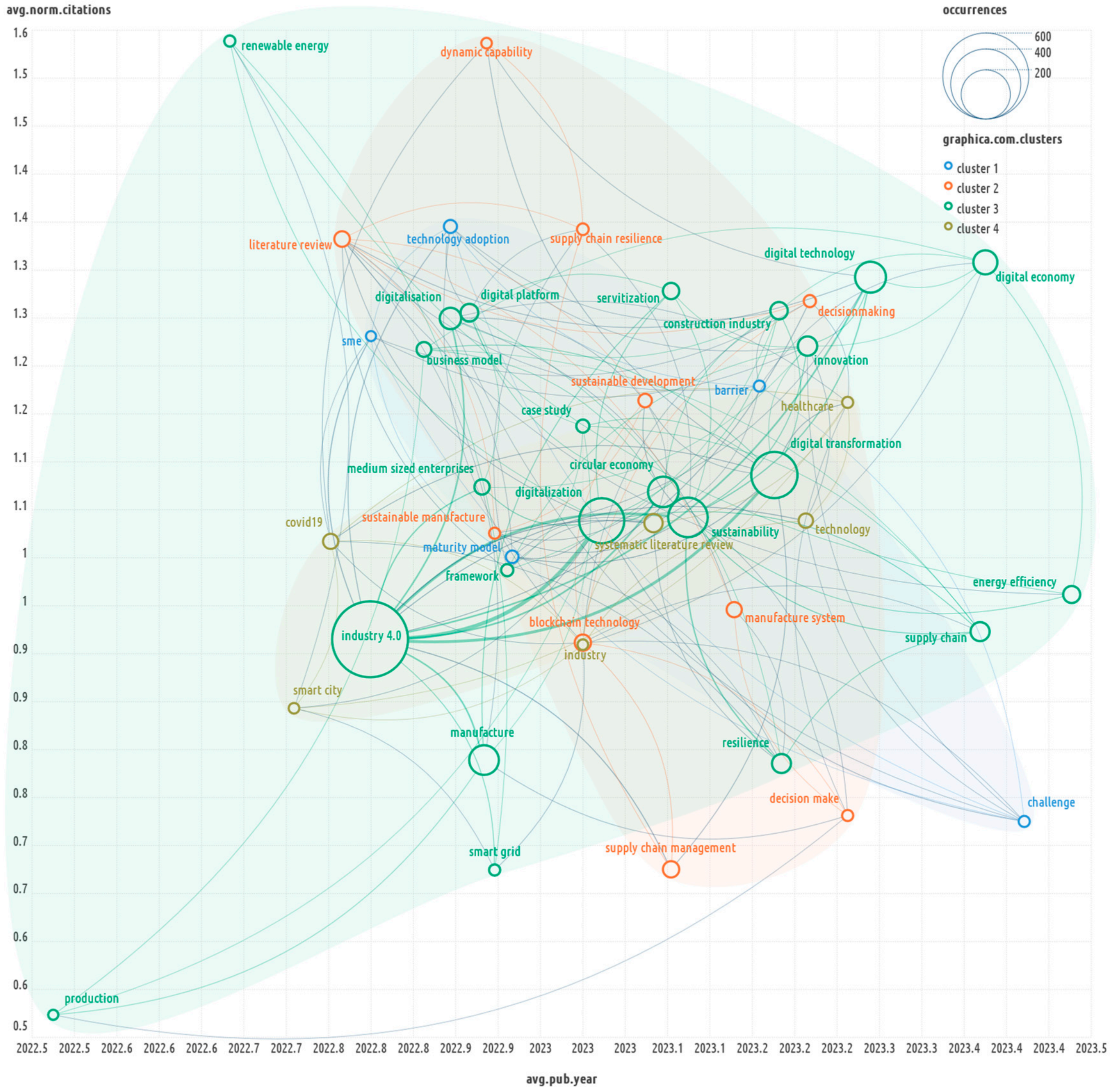


Figure 2.1. Network of Author keywords from the first cluster of Figure 2.

The terms 'renewable energy, digital economy and energy efficiency' may describe relevant research topics, such as those presented in the following papers [13, 14].

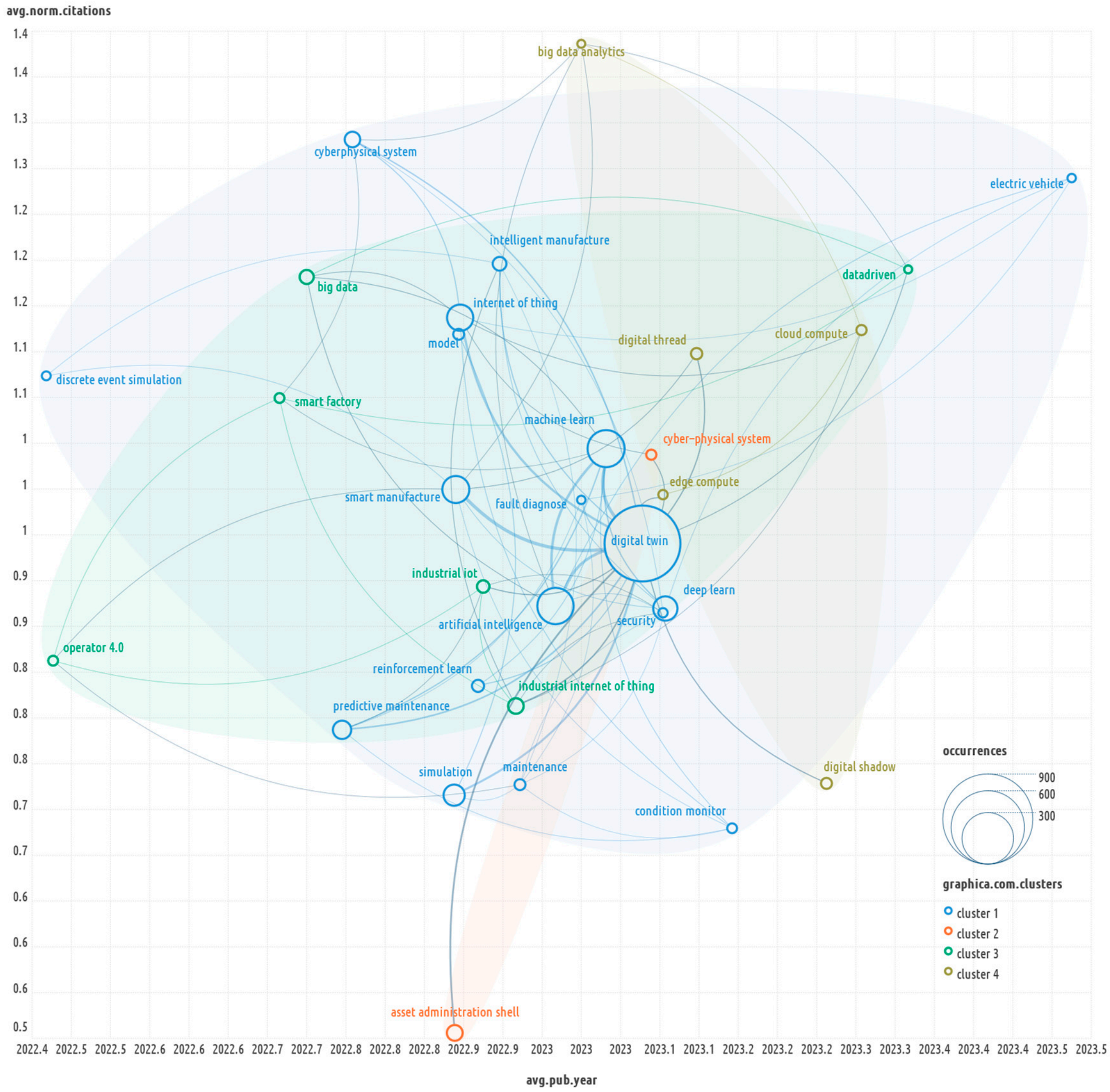


Figure 2.2. Network of Author keywords from the second cluster of Figure 2.

In this case, the topic ‘digital twin’ with the most relevant terms ‘electric vehicle and cyberphysical system’ is best represented. An example of an article revealing this topic is [15], cited 271 times according to ScienceDirect and 397 times according to GOOGLE. Current as of 10–10–2024.

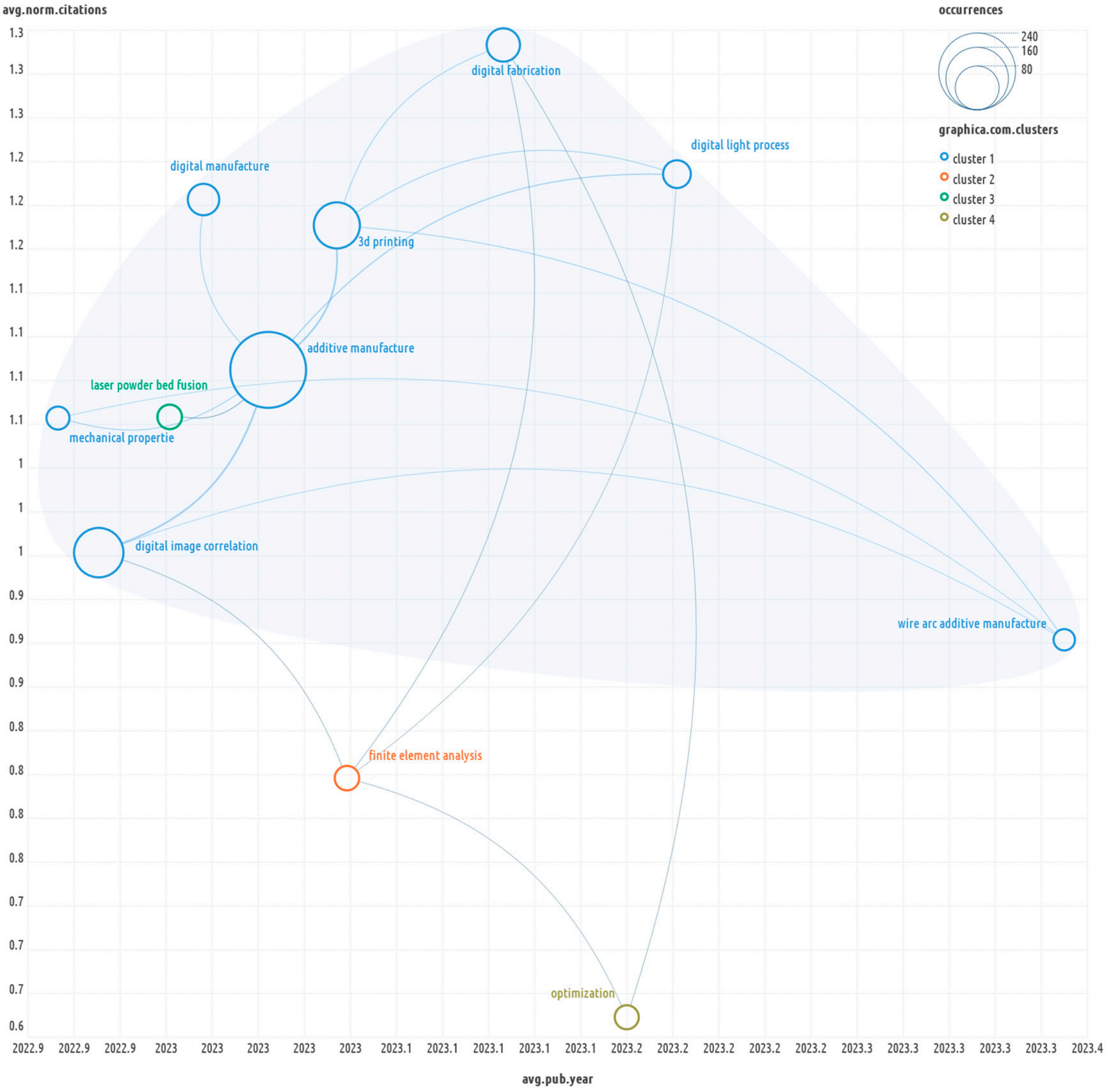


Figure 2.3. Network of Author keywords from the third cluster of Figure 2.

This group of terms is dominated by the topic ‘additive manufacture’. And ‘wire arc manufacture’ is characteristic of new publications, and ‘digital fabrication’ which are more frequently cited. The articles [16, 17] are relevant examples that give the context of the use of these terms.

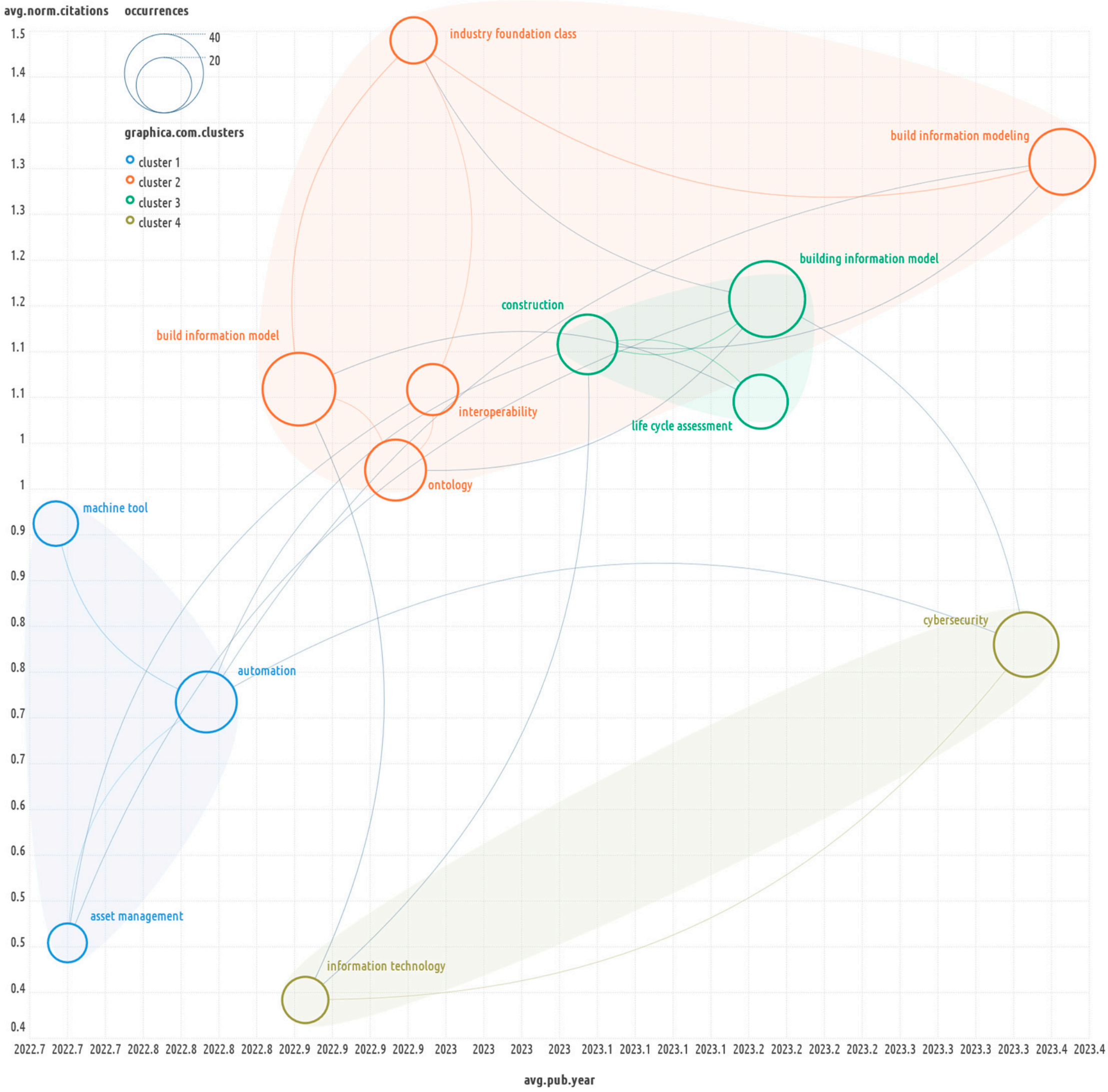


Figure 2.4. Network of Author keywords from the fourth cluster of Figure 2.

Cluster 2 contains not only the maximum number of terms but also the most topical terms ‘build information modeling’ and ‘industry foundation class’ (IFC is a standardized, digital description of the built asset industry).

Context for these terms: “This research explores the application of Building Information Modeling (BIM) in construction and demolition waste (CDW) management and the sustainability of buildings. The method consisted of applying Design Science Research (DSR) to develop a conceptual information model based on the Industry Foundation Classes (IFC)” [18]. “Building Information Modeling (BIM) plays a pivotal role in the construction management of dam engineering projects … while Industry Foundation Classes (IFC) offer a standardized digital representation of the built environment” [19].

This graph clearly shows the importance of keyword normalization: the terms ‘build information modeling’ and ‘build information model’ have a similar meaning, but are placed on different sides of the second cluster. In more general graphs, such as Figure 2, this problem may not be noticeable, but a closer look, Figure 2.4, shows that the lack of normalization significantly affects the clustering results.

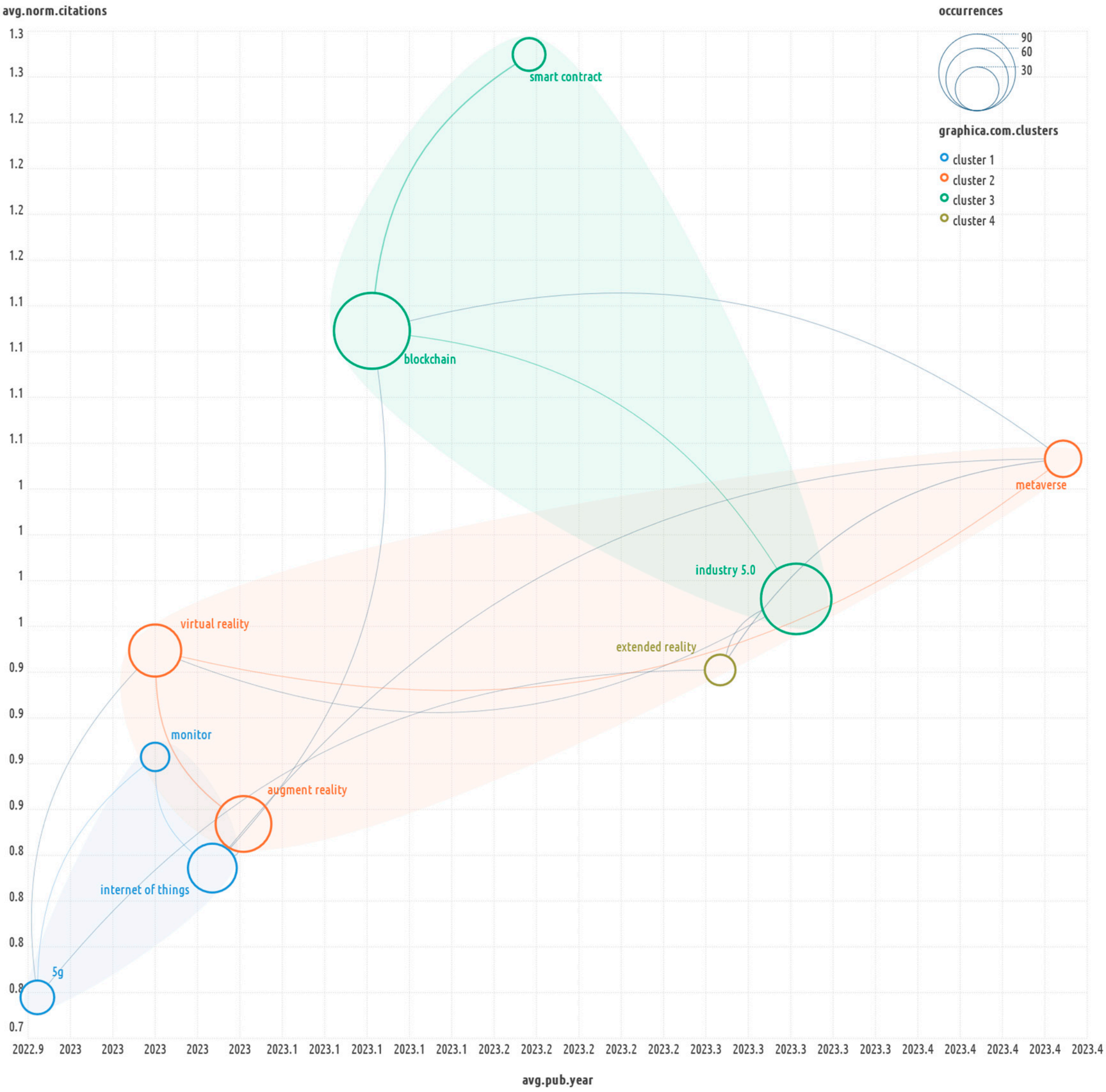


Figure 2.5. Network of Author keywords from the fifth cluster of Figure 2.

‘Smart contract’, ‘blockchain’ and ‘Industry 5.0’ describes the topical theme of this cluster.

Context for the given terms: “the Blockchain concepts and their applications in Marketing through bibliometrics, network, and thematic analyses, which can provide several novel insights … by evaluating the most significant and cited research publications, keywords, institutions, authors’ collaboration network, and finally countries that promote Industry 5.0 (I5.0) businesses” [20]. *For bibliometric research, the topic of using ‘Blockchain concepts’ for ‘Marketing through bibliometrics’ deserves a separate, more in–depth consideration*. [21] — “propose intelligent smart city architecture for vaccine manufacturing with a Smart Contract–based access control model”.

Author Keywords Co-Occurrence Analysis for Journals in Q1

Next, let’s look separately at the distribution of Author keywords for the journals included in Q1, Figure 3.

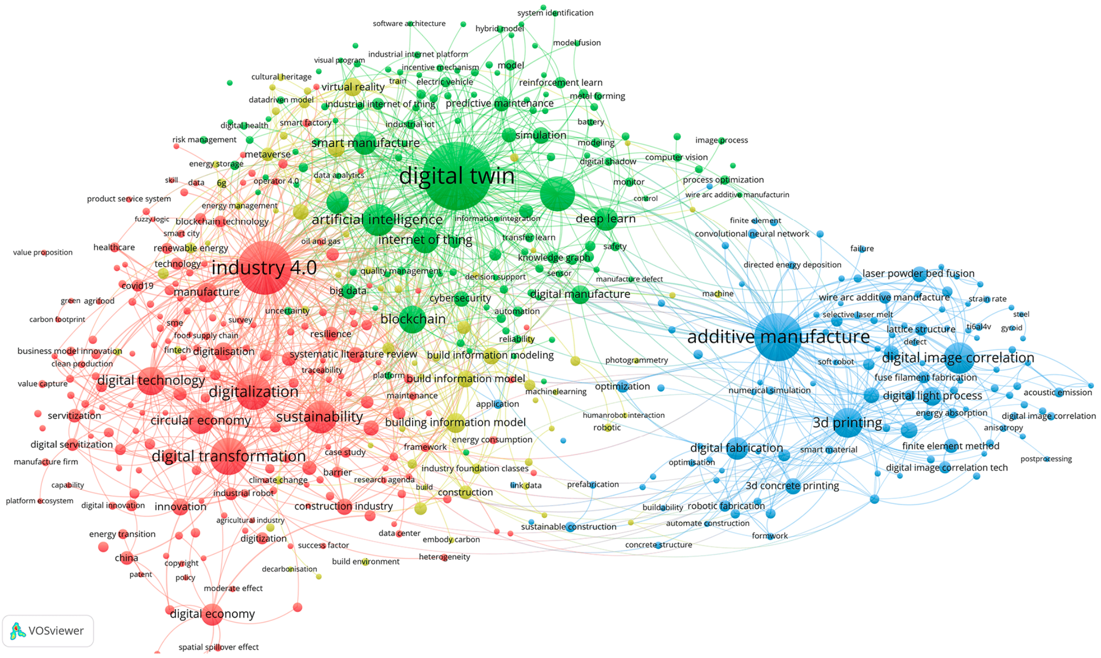


Figure 3. Author keyword co-occurrence network for Q1 journals.

Number of entries for journals included in Q1 — 2698.

Clustering parameters:8263 terms meet 5; 702 meet 3; 500 terms in use; 4 clusters for 50 terms min in cluster.

Table 3 shows 15 Author keywords that are typical for new publications for the journals included in Q1. Obtained by sorting the field score<Avg. pub. year>.

Table 3. Author keywords specific to new publications for journals included in the Q1.

label	cluster	Occurrences	Avg.pub.year	Avg.norm.citations
education	4	4	2024	0.5459
incentive mechanism	2	5	2024	0.9608
nfts	4	3	2024	1.0029
opportunities	1	3	2024	0.7602
reinforce concrete	3	3	2024	0.7602
technology acceptance model	4	4	2024	0.922
modular construction	4	6	2023.833	1.1539
explainable ai	2	5	2023.8	0.9477
industrial metaverse	2	5	2023.8	1.4024
text analysis	4	5	2023.8	1.4442
cementitious composite	3	4	2023.75	0.7949
green supply chain management	1	4	2023.75	1.186
humancentric	2	4	2023.75	1.3667
intelligent system	4	4	2023.75	1.1007
moderate effect	1	4	2023.75	1.1129

NFTs (or "non–fungible tokens") are a special kind of cryptoasset in which each token is unique

The terms 'Incentive mechanism', NFTs, 'technology acceptance model' reflect the current topic according to this table.

The context for these terms is disclosed in the articles: [22] — "The model captures the core categories that affect trust toward crypto–tokens applications ... to develop strategies that ... create value with crypto–tokens and Web3.0 economies". "Actually, the Technology acceptance model (TAM) was initially formulated by Davis to explore the relationships among the elements like "perceived usefulness" (PU), "perceived ease of use" (PEOU), Attitude (ATT), Behavioural Intention (BI) and actual use (AU)" [23].

Table 4. Author keywords specific to publications of Q1 journals with high average normalized citation rate.

label	cluster	Occurrences	Avg.pub.year	Avg.norm.citations
data	1	4	2023.25	2.245
battery	2	5	2023.6	2.1968
information fusion	2	4	2023.25	2.0612
sensor	2	5	2022.8	1.8612
energy transition	1	9	2023.222	1.6502
digital concrete	3	8	2022.875	1.567
financial performance	1	4	2022.75	1.5522
remain useful life	2	3	2023.333	1.5301
electronic	2	4	2022.75	1.5294
evaluation	1	3	2022.333	1.5209
federate learn	2	7	2023.714	1.4955
process plan	2	3	2023	1.4939
ambidexterity	1	3	2022.333	1.4895
patent	1	3	2022.667	1.4493
text analysis	4	5	2023.8	1.4442

The terms 'data', 'battery', 'information fusion' define the relevance topic from this table. The title of the paper itself [24] reveals the context of these terms – 'Multi sensor fusion methods for state of charge estimation of smart lithium–ion batteries'.

Diagrams Created with the Scimago Graphica

Figures 3x1–4 are plotted using the Scimago Graphica program for each of the 4 clusters shown in Figure 3 in a similar way to Figures 2x1–5.

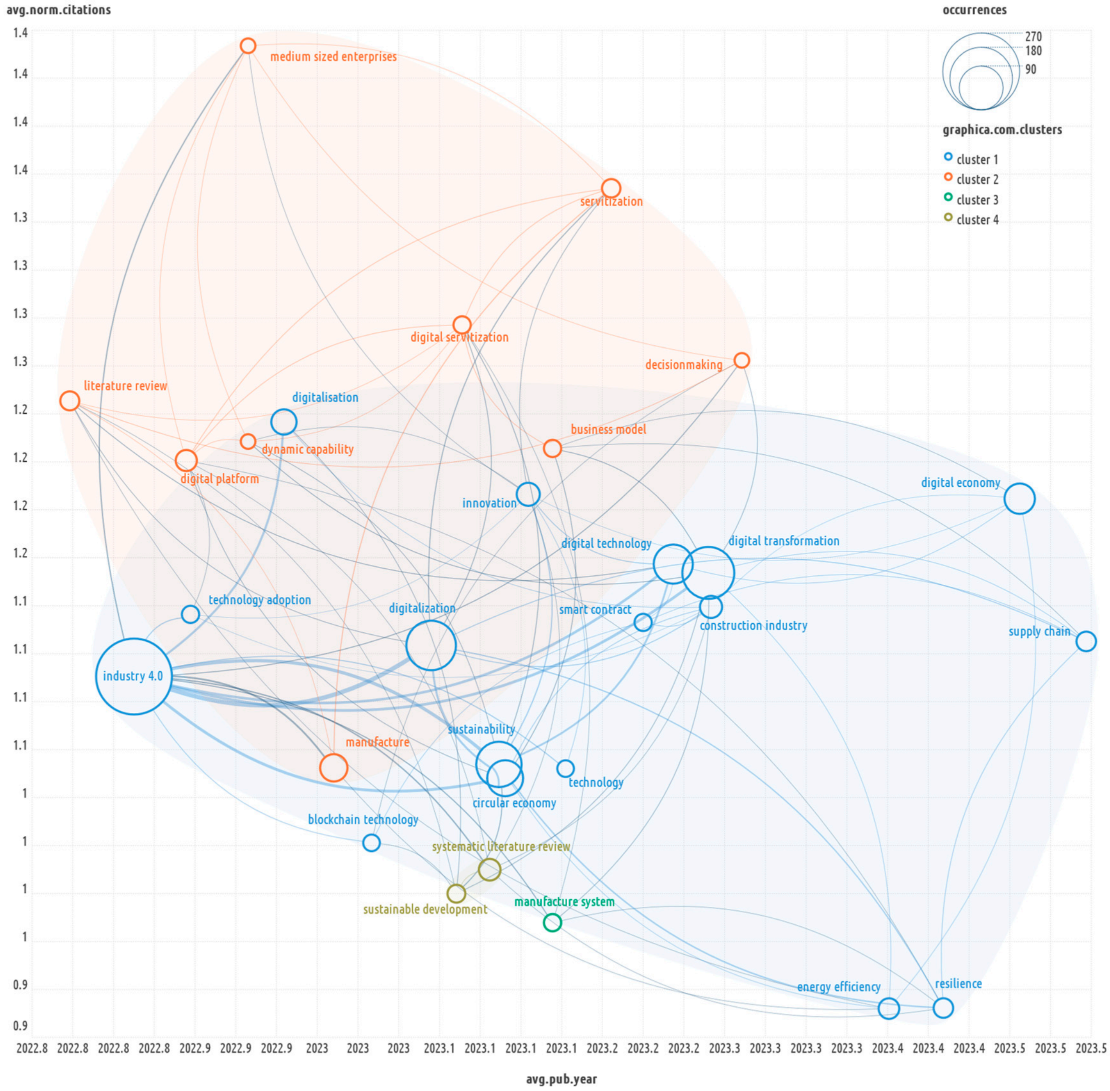


Figure 3.1. Network of Author keywords from the first cluster of Figure 3.

Digital economy, especially 'digital economy and supply chain', are among the largest clusters of emerging topics in Q1 journal publications.

An example of a publication reflecting this theme is the article [25] showed that a study of Chinese public companies from 2007 to 2020 found a positive correlation between the level of digital transformation of large customers and the level of digital transformation of their suppliers, especially if the customers are important to the suppliers.

The authors of [26] have shown that the "The result shows that "absence of urgency for supply chain digitalization", lack of proper innovation strategies", and "Inadequate leadership to lead digital transformation "are the highest–ranked digital supply chain barriers that needs to be overcome on a priority basis".

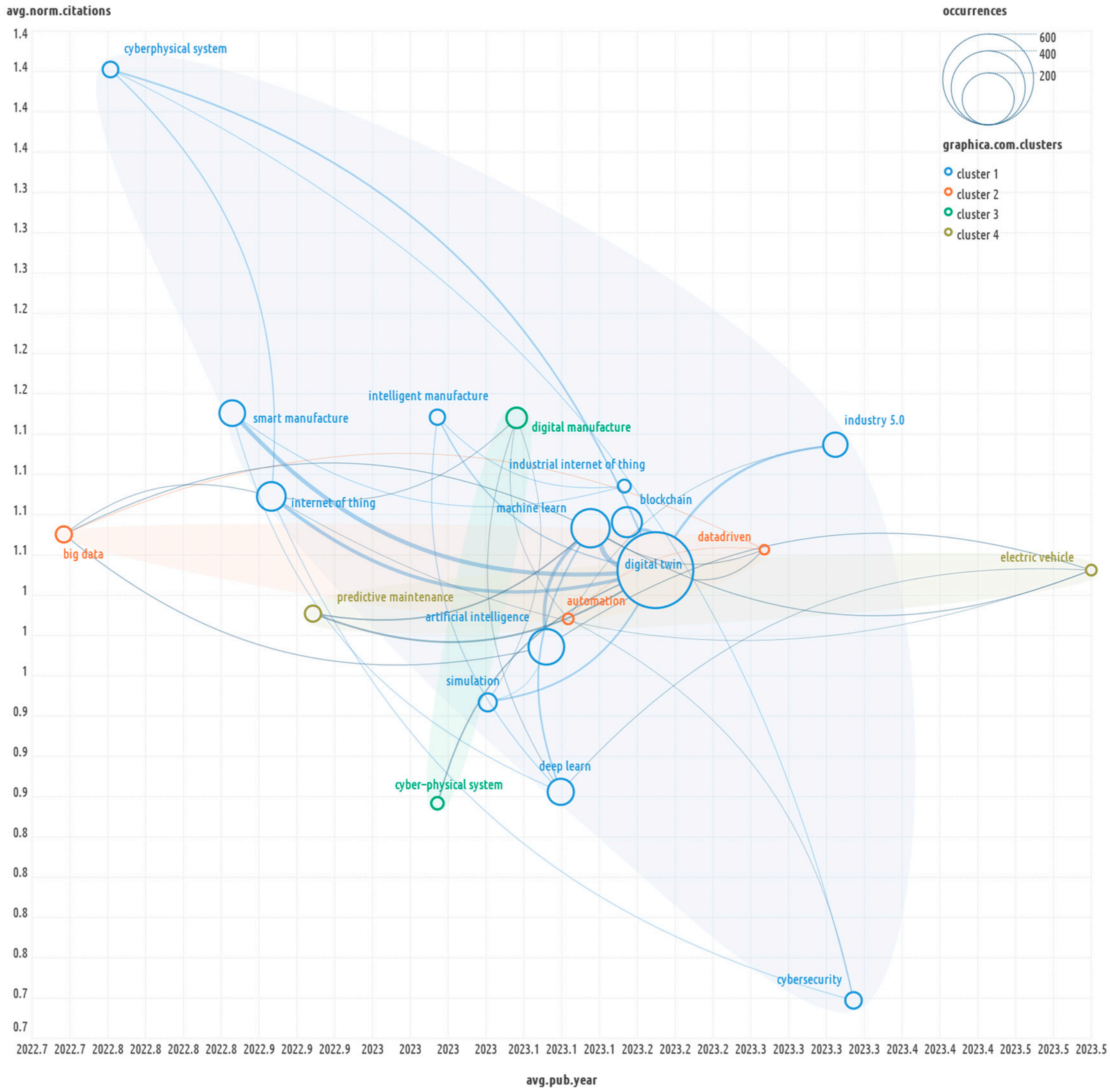


Figure 3.2. Network of Author keywords from the second cluster of Figure 3.

'Cyberphysical systems' and 'Industry 5.0' are the topics of interest in this cluster.

Context for these terms: "… transforming industrial automation into industry 5.0 compliances to attain full autonomy with minimal human intervention … Smart Cyber–Physical System (SCPS) is the most important part of the fourth industrial revolution, where diffident programmed embedded systems are networked together to perform, computation, communication, control, and actuation" [27]. Further details on this topic can be found in Special issue. Cyber–Physical System for Autonomous Process Control in Industry 5.0[4].

[4] https://www.sciencedirect.com/special-issue/102XPP5FG52 — Cyber-Physical System for Autonomous Process Control in Industry 5.0, Last update 16 February 2024

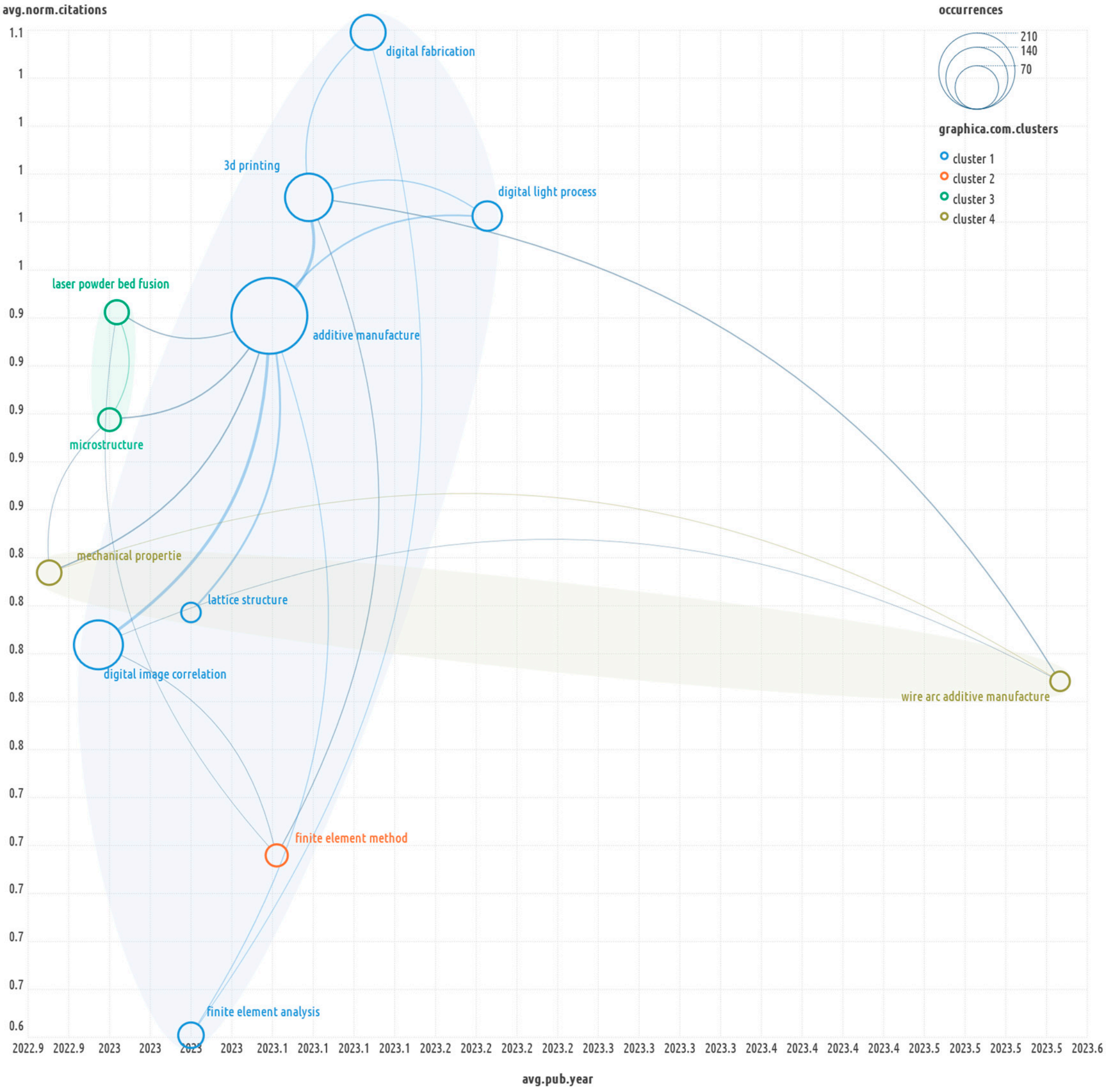


Figure 3.3. Network of Author keywords from the third cluster of Figure 3.

The theme: ‘digital fabrication’, ‘digital light process’ is the most cited topic in this cluster and is close to the similar theme presented in Figure 2.3. At the same time, ‘wire arc manufacture’ was separated into a different topic. This example shows how clustering depends on the chosen data slice, emphasizing that filters are part of the question we want to answer.

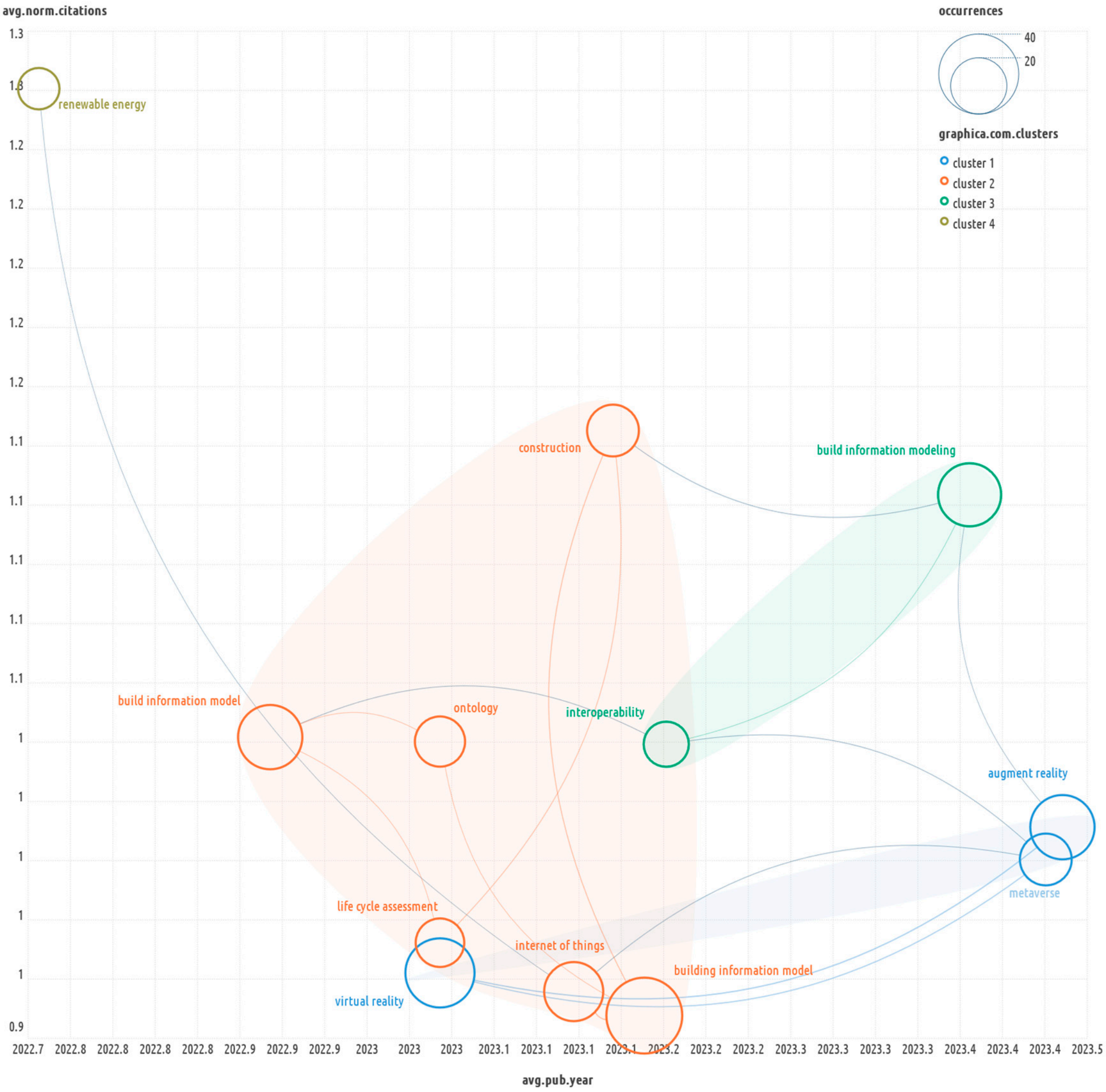


Figure 3.4. Network of Author keywords from the fourth cluster of Figure 3.

It can be seen here that the topic ‘augment realty’ appears frequently in new publications, but does not have a high citation rate. The topic ‘build information models’ has already appeared in Figure 2.4 under the spellings ‘build information modeling’ and ‘build information model’, which again shows the importance of the keyword preparation stage in cluster analysis.

Author Keywords Co-Occurrence Analysis for Journals Not in Q1

The co-occurrence clustering of Author keywords shown in Figure 4 is constructed on the basis of 968 records with the following parameters: 2919 of all terms, 100 meet 5 and more times, 230 meet 3 and more times, give 4 clusters for 30 min terms in cluster.

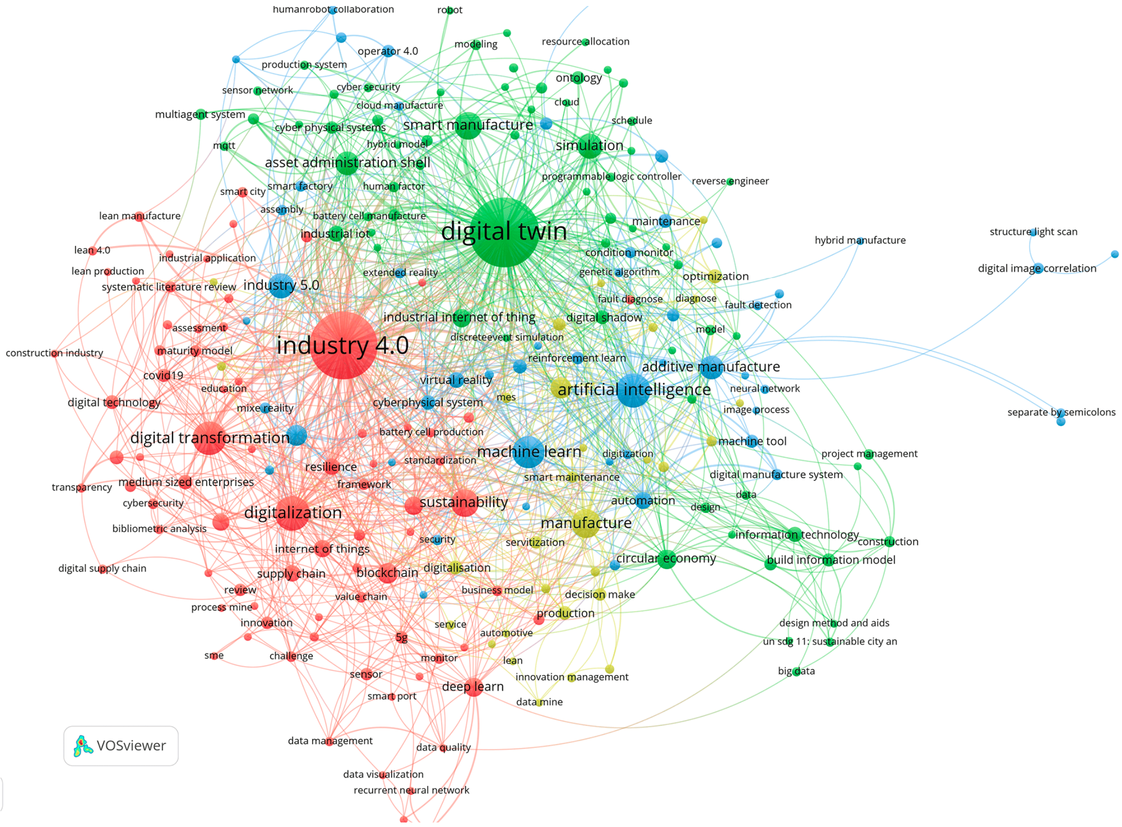


Figure 4. Author keyword co-occurrence network for non–Q1 journals.

The total number of publications not included in Q1 was less than those included: 2698 and 968, respectively. Therefore, the 'total link strength' filter was not used for the Author keywords of such publications when constructing the 4x1–4 graphs. Taking into account that the purpose of this paper was not to analyze the topic of digital industry in detail, but to demonstrate the possibilities of the proposed research approach, not using this filter allowed us to show the reader how graphs reflecting a larger number of terms can look like. Only the 'cluster' filter was used.

Table 5. Author keywords specific to new publications for journals not included in the Q1.

label	cluster	Occurrences	Avg.pub.year	Avg.norm.citations
strategic management	1	3	2024	1.2048
energy efficiency	1	5	2023.8	1.07
digital product passport	2	4	2023.75	1.1367
aas	2	3	2023.667	0.7873
digital technology	1	9	2023.667	1.0157
education	1	3	2023.667	0.9133
healthcare	1	3	2023.667	1.244
hybrid model	2	3	2023.667	0.9881
life cycle assessment	2	3	2023.667	0.9291
logistic	3	3	2023.667	0.9842
product development	1	3	2023.667	0.9842

transfer learn	4	3	2023.667	1.0471
modeling	2	5	2023.6	0.94
innovation	1	7	2023.571	0.9785
cloud manufacture	3	4	2023.5	0.9968

The terms ‘strategic management’ and ‘energy efficiency’ are more frequent in new publications and have high citation rates.

Table 6. Author keywords specific to publications of not in Q1 journals with high average normalized citation rate.

label	cluster	Occurrences	Avg.pub.year	Avg.norm.citations
sensor network	2	4	2022	1.5679
financial technology	3	3	2022.333	1.2948
healthcare	1	3	2023.667	1.244
systematic literature review	1	6	2023	1.2355
smart city	1	5	2022.6	1.2162
assembly	3	5	2022.2	1.2066
strategic management	1	3	2024	1.2048
smart grid	1	6	2022.667	1.2008
data mine	4	3	2022.667	1.1871
sustainable manufacture	1	3	2022.667	1.1855
digital economy	1	6	2022.667	1.1838
industry	1	4	2022.75	1.1755
biologicalisation	1	3	2023	1.1714
hybrid manufacture	3	3	2023	1.1714
lean production	1	3	2022.333	1.1618

The terms ‘sensor network’ and ‘financial technology’ are found in earlier papers, but have a good citation rate.

Diagrams Created with the Scimago Graphica

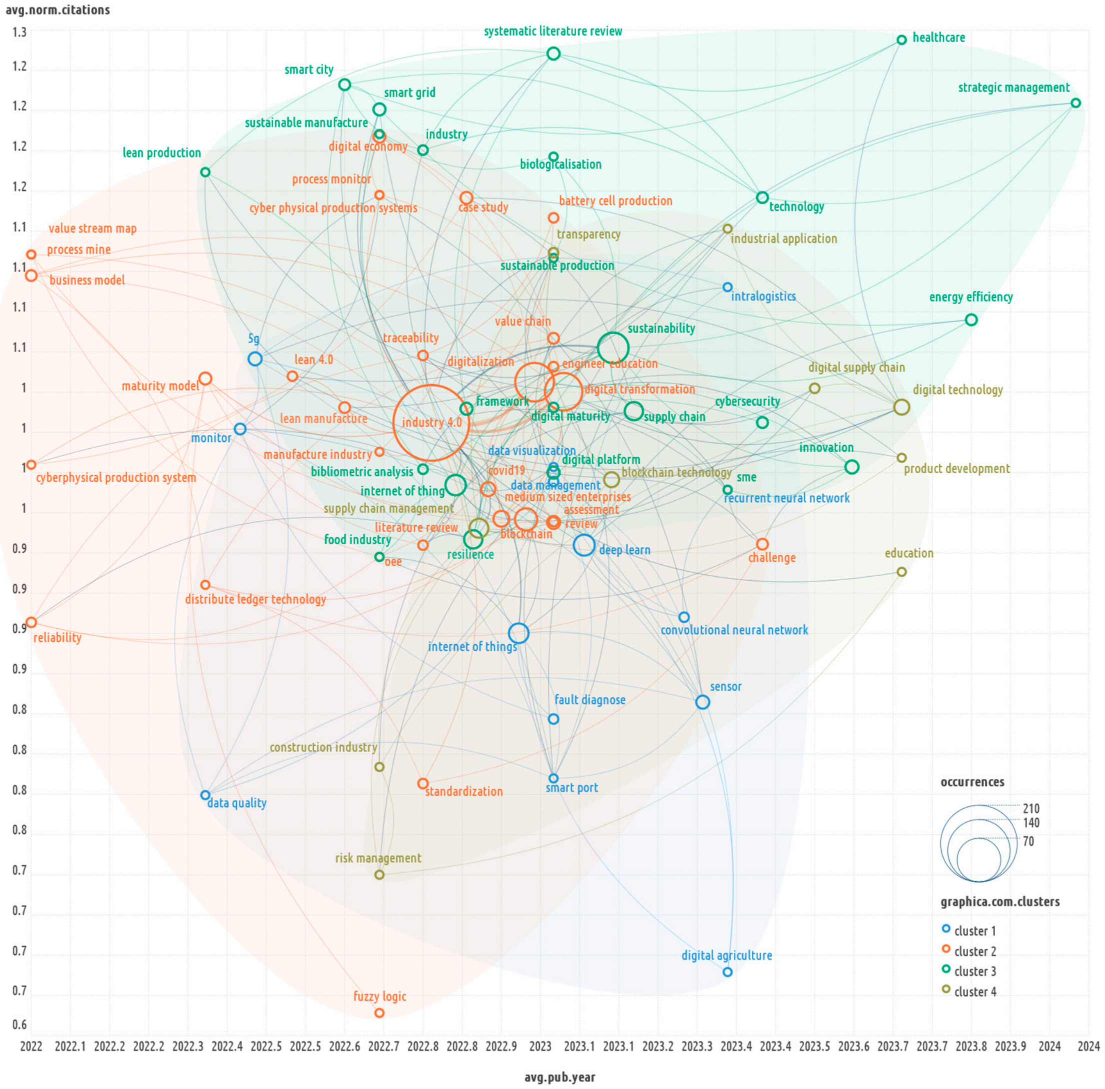


Figure 4. 1. Network of Author keywords from the first cluster of Figure 4.

‘Systematic literature review’ and ‘strategic management’ can be an interesting topic. In [28], which is a systematic literature review, the term ‘strategic management’ appears extensively in the form of references to the journal: Analysis & Strategic Management.

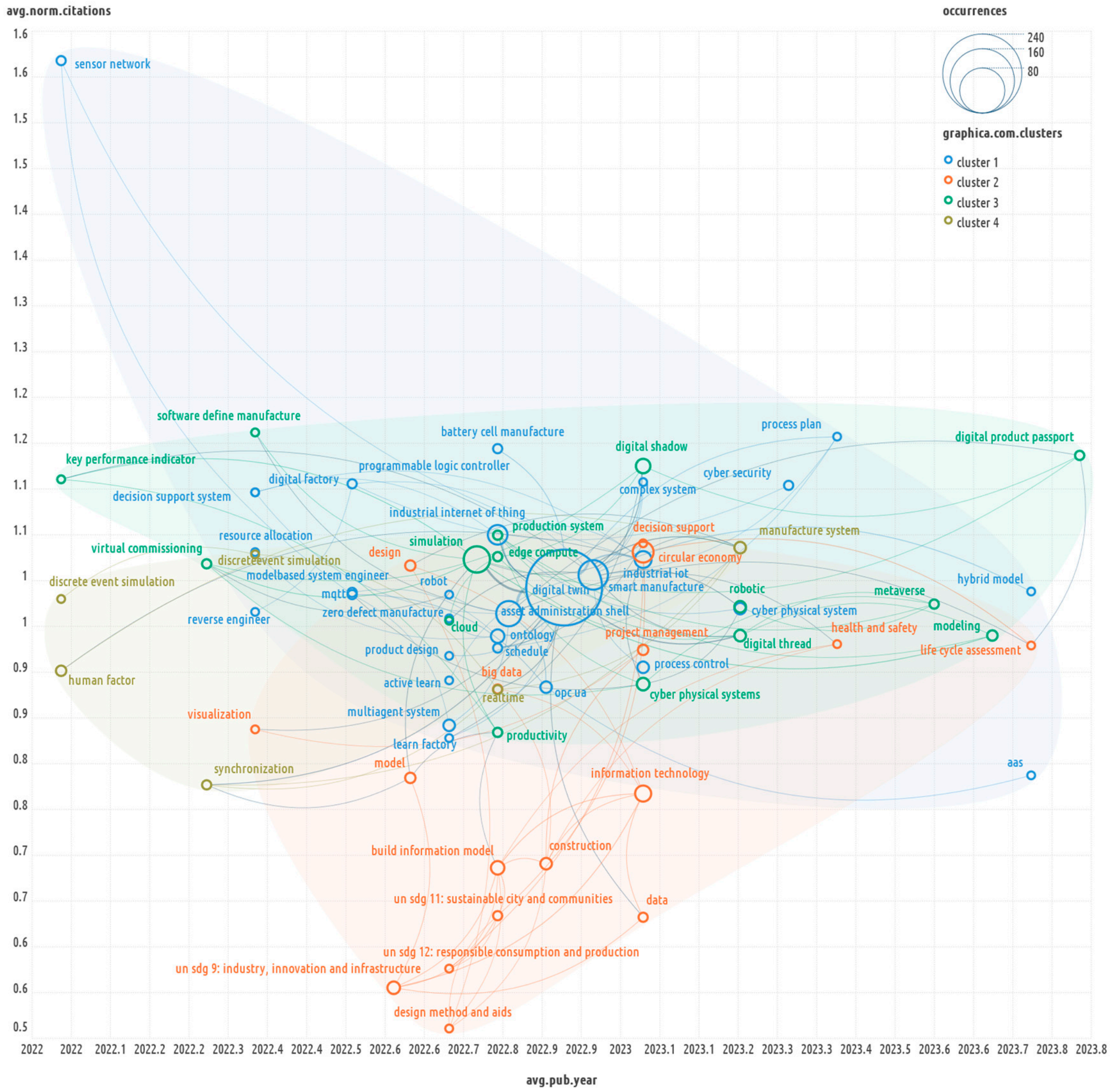


Figure 4.2. Network of Author keywords from the second cluster of Figure 4.

In the topic of ‘sensor network’ and ‘aas’, the former term is highly cited and the latter term is more common in new publications. Context: “Asset Administration Shell (AAS) – a machine–readable representation of an asset in a digitalized production system” … “AAS can provide static information, such as the manufacturer, type or serial number, as well as dynamic data such as the current reading of a sensor” [29].

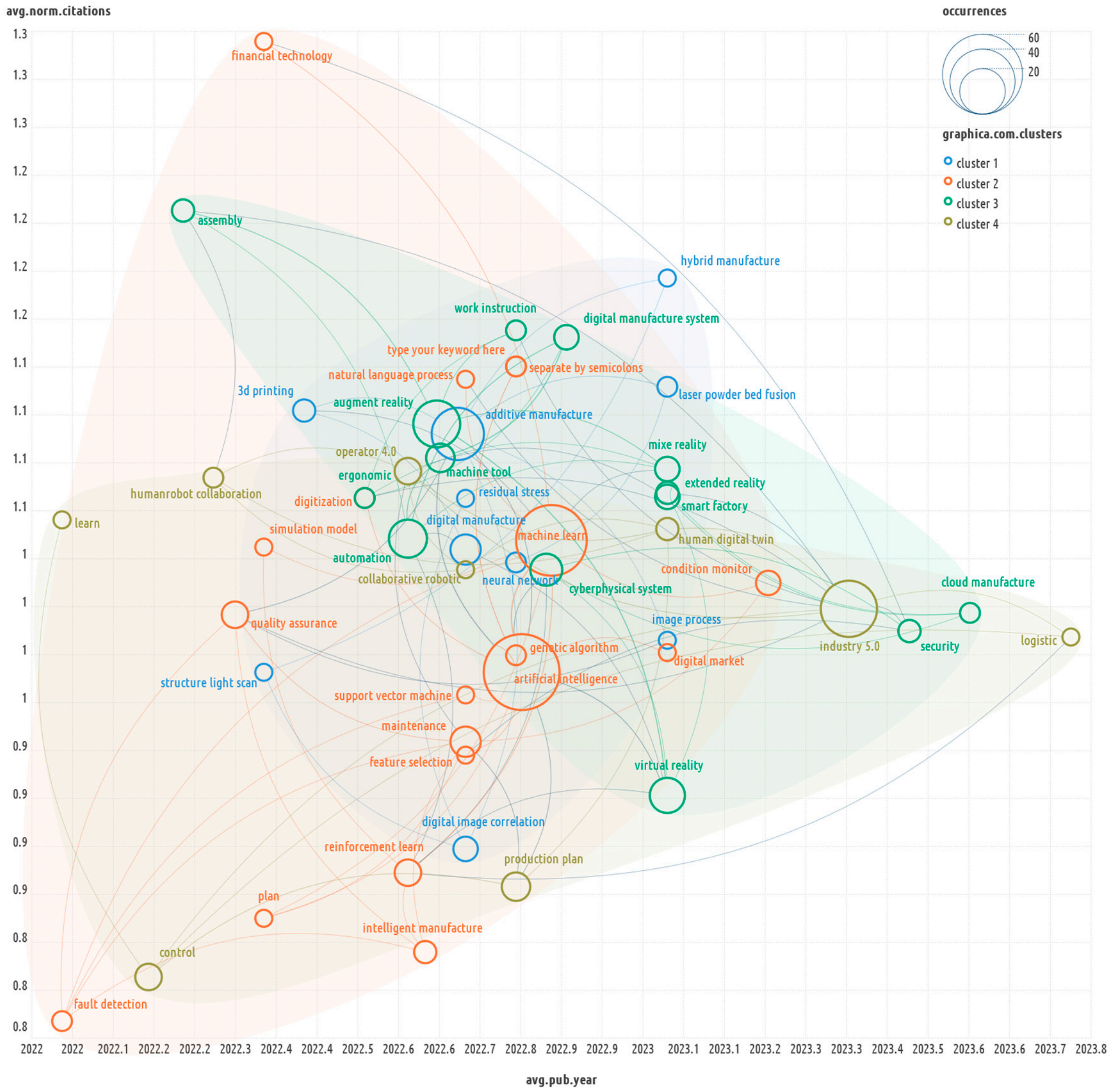


Figure 4.3. Network of Author keywords from the third cluster of Figure 4.

The second, red cluster in this figure refers to 'machine leaning' and 'artificial intelligence', with the term 'financial technology' appearing in articles with higher citations. Machine leaning algorithms are used in analyzing issues related to cryptocurrencies, an example of such work is the following [30].

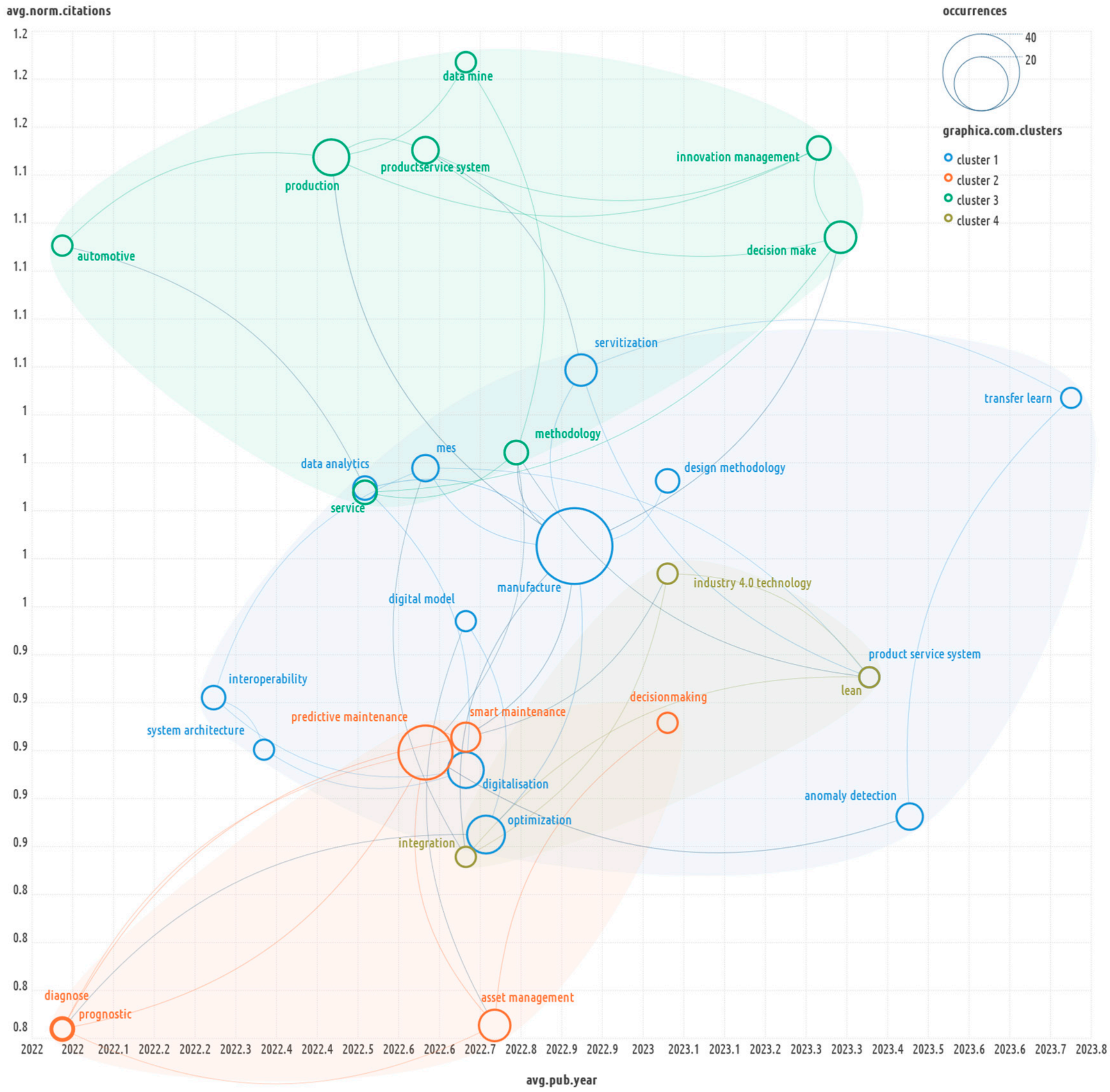


Figure 4.4. Network of Author keywords from the fourth cluster of Figure 4.

The topic of the fourth cluster: ‘data mining’, ‘innovation management’ and ‘decision making’ is well covered in the works [31, 32].

Diagrams Created with the Scimago Graphica

Previously, when building a keyword network with Scimago Graphica, filters were used to set the threshold for the ‘total link strength’ and ‘cluster’ parameters. If the task is to analyze the topics of articles published in the most cited journals, it is advisable to introduce filtering by the ‘avg.norm.citation’ parameter. Figures 3.1cit–3.3–4cit show the results of Author keyword clustering for publications published in journals included in Q1, but unlike Figures 3.1–3.4 filtering was used: ‘total link strength’>20 and ‘avg.norm.citation’>1,1.

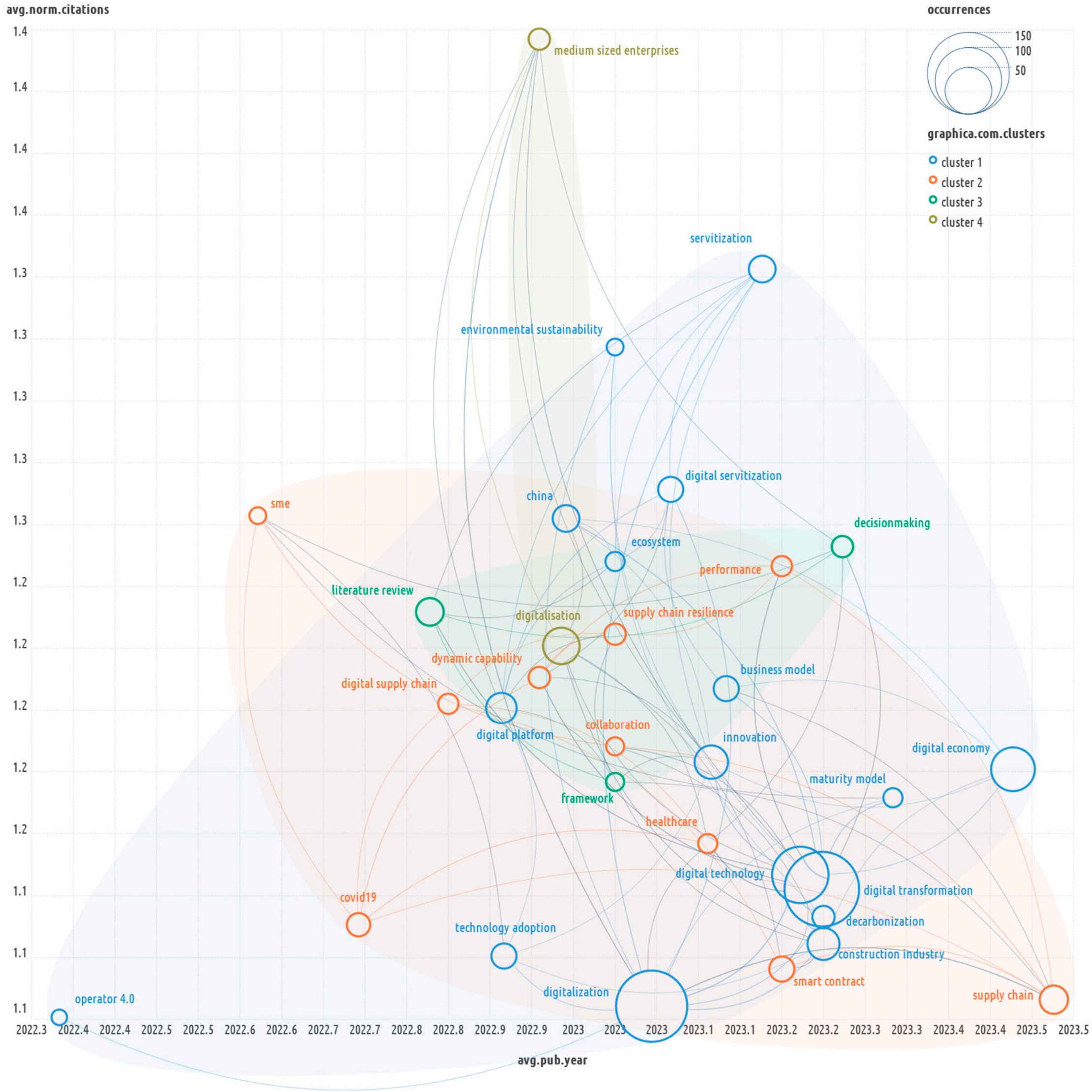


Figure 3.1cit. Network of Author keywords from the first cluster of Figure 3 when filtering 'total link strength'>20 and 'avg.norm.citation'>1,1.

The topic 'servitization' and 'digital economy' can be seen as a contender for relevance. The publication [33] provides a context for these two terms in its title. Another example from this paper is the line: "the impact of the digital economy on the servitization of industrial structures in the eastern, central, and western regions".

Definition: Servitization is a shift in manufacturing that enables the creation of innovative business models and revenue streams by incorporating value–added services into their products.

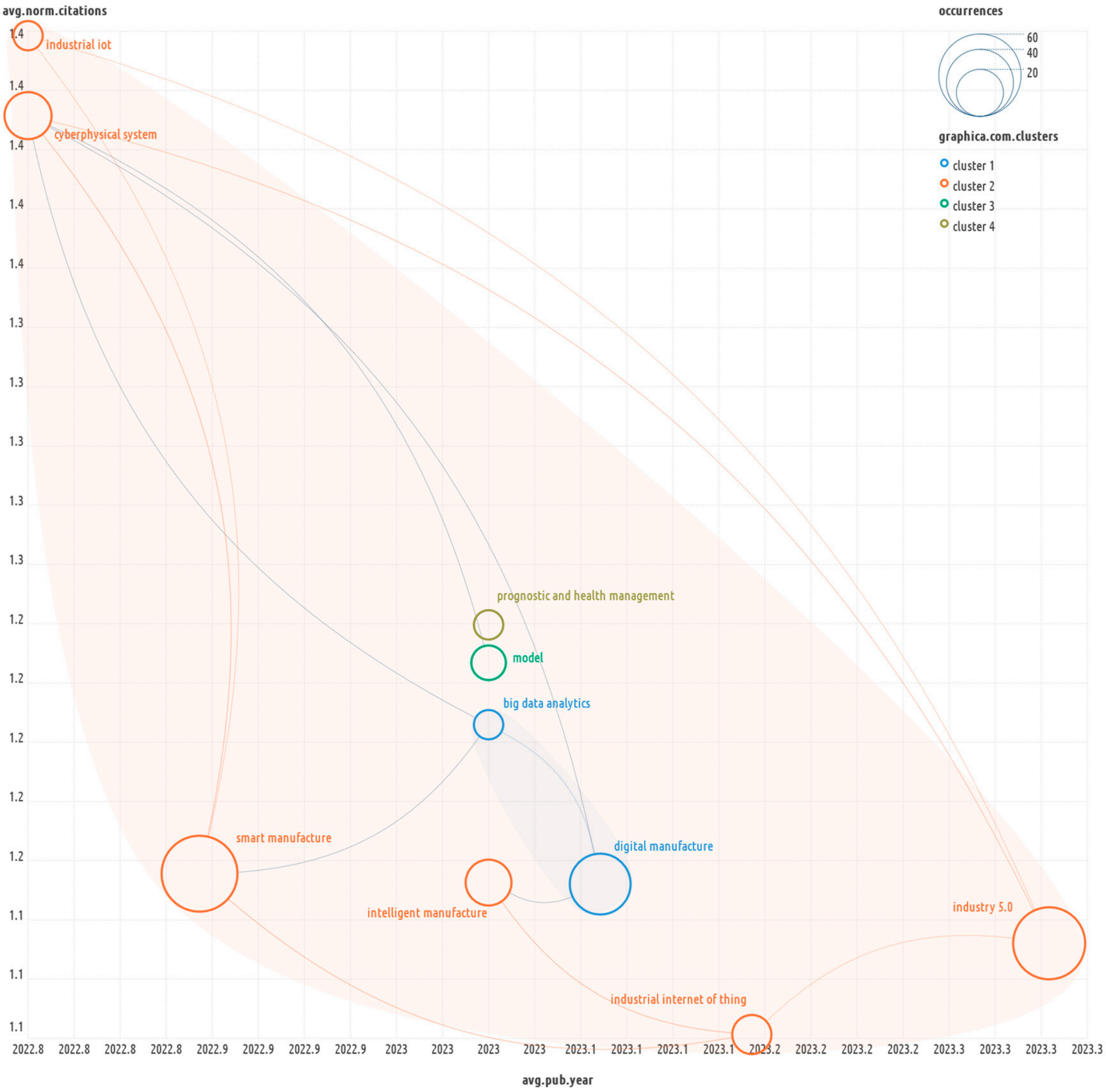


Figure 3.2cit. Network of Author keywords from the second cluster of Figure 3 when filtering 'total link strength'>20 and 'avg.norm.citation'>1,1.

The theme of 'industrial iot' and 'industry 5.0' dominates in this chart. Strings "Emerging technologies such as blockchain and digital twins are essential for … Industry 5.0 … the growing number of industrial IoT nodes … enable secure transmission … difficult" from [34] reflect the context of the terms.

The keyword 'industrial iot' is an example of how it is not enough to simply replace individual keywords as abbreviations with their full names, it is also necessary to replace abbreviations in complex keywords, for example using the sed utility.

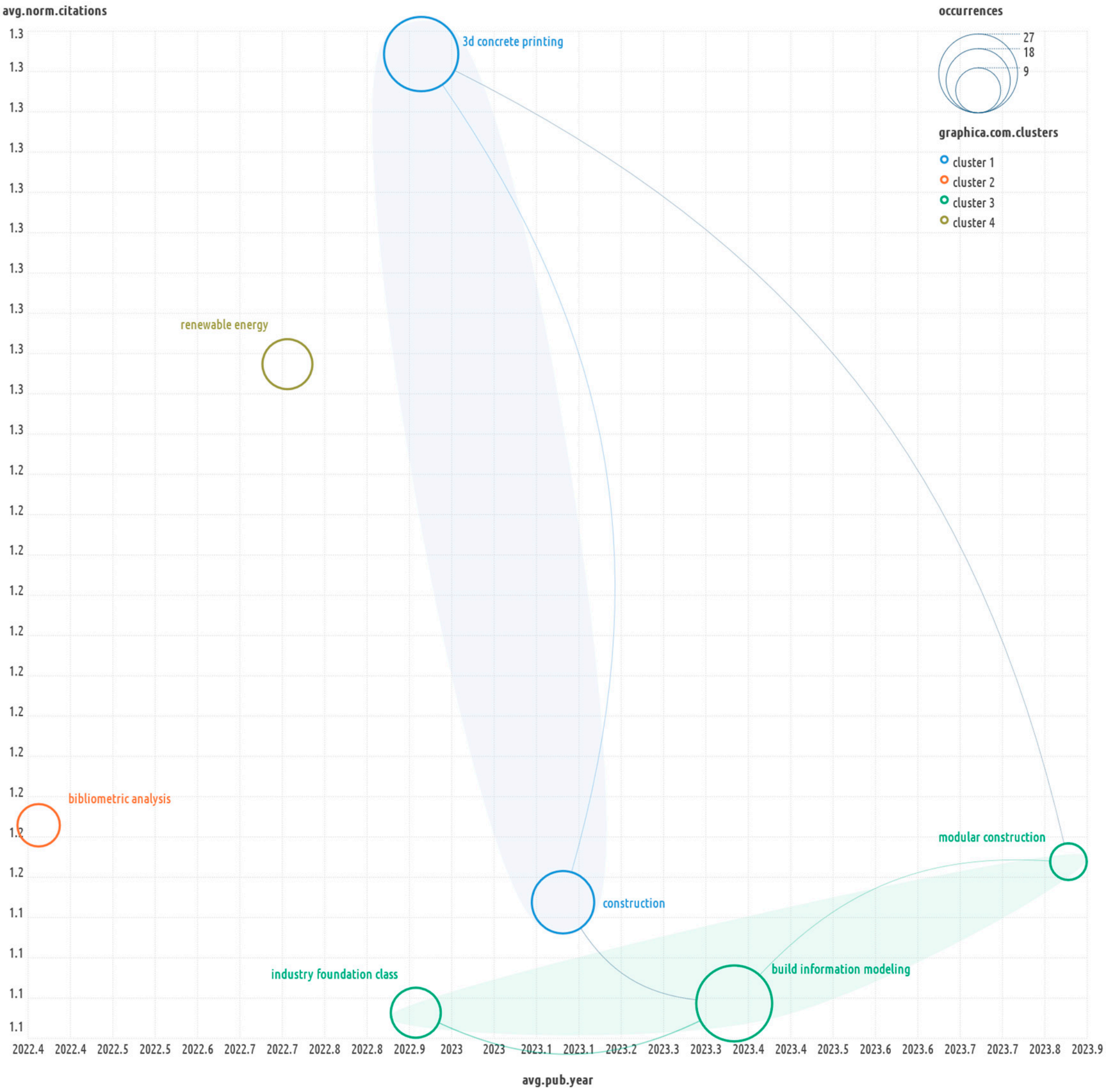


Figure 3.3–4cit. Network of Author keywords from the third and fourth cluster of Figure 3 when filtering 'total link strength'>20 and 'avg.norm.citation'>1,1.

The three figures above show that high cited keywords form a clearly dominant cluster, as shown in the Figure 3.1cit.

The topic '3d concrete printing' can be found in the review publications for 2024 [35, 36, 37].

Conclusions

An assessment of the importance of keywords in new publications (which have not yet been cited) is suggested by the average citation of articles in the journal over the last two years, values of which are available, for example, on the Scimago Journal & Country Rank platform.

Since the paper is in the style of a proof of concept, a feature like Journal Hirsch can also be used to assess the potential relevance of keywords.

The importance of pre-processing of Author keywords before the procedure of their clustering is shown. It should include at least such steps of their normalization as the use or absence of hyphens or short dashes in complex words, the conversion abbreviations to full names, lemmatization of terms and even manual checking. For example, there are different spellings of the term 'decision making', 'decision–making' and 'decisionmaking'. The author of this paper prefers lemmatization to stemming, as the latter can make terms difficult to read graphs.

A sequential use of the programs VOSviewer and Scimago Graphica is proposed. The former is used to present an overall picture of the research landscape, while the latter is used for a more detailed analysis of individual slices of bibliometric data, including those obtained from the VOSviewer program.

In order to assess the relevance of keywords, it is proposed to display the network of their co-occurrence in the coordinates 'Avg.pub.year' vs. 'Avg.norm.citation' instead of the usual layout. The use of the 'convex hull' facilitates the perception of cluster boundaries and thus the selection of topical terms in the cluster. The Scimago Graphica program offers this possibility.

The use of filters in Scimago Graphica greatly facilitates the creation of the slices analysed by the data and is recommended for their use.

After the graphical analysis of the data and the highlighting of the terms, the author of this paper considers it appropriate to offer the context of their appearance in the form of citation strings from publications that reflect well the context of the use of the terms, as well as providing definitions of not all known terms.

As for the topic of industry digitalization itself, it is not only a technical and technological issue, but also an economic one, reflected in terms such as the digital economy and Industry 5.0.

References

1. Van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010;84:523–38. https://doi.org/10.1007/s11192–009–0146–3.
2. Neylon C, Wu S. Article–Level Metrics and the Evolution of Scientific Impact. PLoS Biol 2009;7:e1000242. https://doi.org/10.1371/journal.pbio.1000242.
3. Eyre–Walker A, Stoletzki N. The Assessment of Science: The Relative Merits of Post–Publication Review, the Impact Factor, and the Number of Citations. PLoS Biol 2013;11:e1001675. https://doi.org/10.1371/journal.pbio.1001675.
4. McEvoy NL, Latour JM. From impact factors to Altmetrics: What numbers are important in publishing your paper? Nursing in Critical Care 2023;28:4–6. https://doi.org/10.1111/nicc.12925.
5. DiBartola SP, Hinchcliff KW. Metrics and the Scientific Literature: Deciding What to Read. Veterinary Internal Medicne 2017;31:629–32. https://doi.org/10.1111/jvim.14732.
6. Meng F, Zhou K, Bu Y, Huang W–B, Zhang P, Long F, et al. Keywords Extraction and Thesaurus Construction for Domain News. Procedia Computer Science 2022;214:837–44. https://doi.org/10.1016/j.procs.2022.11.249.
7. Павлова ИА. ПОСТРОЕНИЕ КАРТЫ СОПРИСУТСТВИЯ КЛЮЧЕВЫХ СЛОВ ПО ТЕМЕ «КАПИТАЛ ЗДОРОВЬЯ» В ПРОГРАММЕ VOSVIEWER. Jwt 2023;49:38–54. In Russian. https://doi.org/10.18799/26584956/2023/2/1592.
8. Hamdan W, Alsuqaih H. Research Output, Key Topics, and Trends in Productivity, Visibility, and Collaboration in Social Sciences Research on COVID–19: A Scientometric Analysis and Visualization. Sage Open 2024;14:21582440241286217. https://doi.org/10.1177/21582440241286217.
9. 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.
10. Hassan–Montero Y, De–Moya–Anegón F, Guerrero–Bote VP. SCImago Graphica: a new tool for exploring and visually communicating data. EPI 2022:e310502. https://doi.org/10.3145/epi.2022.sep.02.
11. Wang Y, Shi J, Qu G. Research on collaborative innovation cooperation strategies of manufacturing digital ecosystem from the perspective of multiple stakeholders. Computers & Industrial Engineering 2024;190:110003. https://doi.org/10.1016/j.cie.2024.110003.
12. Neef T, Müller S, Mechtcherine V. Integrating continuous mineral–impregnated carbon fibers into digital fabrication with concrete. Materials & Design 2024;239:112794. https://doi.org/10.1016/j.matdes.2024.112794.
13. Zheng M, Wong CY. The impact of digital economy on renewable energy development in China. Innovation and Green Development 2024;3:100094. https://doi.org/10.1016/j.igd.2023.100094.
14. Yi J, Dai S, Li L, Cheng J. How does digital economy development affect renewable energy innovation? Renewable and Sustainable Energy Reviews 2024;192:114221. https://doi.org/10.1016/j.rser.2023.114221.

15. Bhatti G, Mohan H, Raja Singh R. Towards the future of smart electric vehicles: Digital twin technology. Renewable and Sustainable Energy Reviews 2021;141:110801. https://doi.org/10.1016/j.rser.2021.110801.
16. Kumar N, Bhavsar H, Mahesh PVS, Srivastava AK, Bora BJ, Saxena A, et al. Wire Arc Additive Manufacturing – A revolutionary method in additive manufacturing. Materials Chemistry and Physics 2022;285:126144. https://doi.org/10.1016/j.matchemphys.2022.126144.
17. Li H, Shi X, Wu B, Corradi DR, Pan Z, Li H. Wire arc additive manufacturing: A review on digital twinning and visualization process. Journal of Manufacturing Processes 2024;116:293–305. https://doi.org/10.1016/j.jmapro.2024.03.001.
18. Schamne AN, Nagalli A, Soeiro AAV, Poças Martins JPDS. BIM in construction waste management: A conceptual model based on the industry foundation classes standard. Automation in Construction 2024;159:105283. https://doi.org/10.1016/j.autcon.2024.105283.
19. Zhang S, Zhang S, Wang C, Zhu G, Liu H, Wang X. Extended IFC–based information exchange for construction management of roller–compacted concrete dam. Automation in Construction 2024;163:105427. https://doi.org/10.1016/j.autcon.2024.105427.
20. Nikseresht A, Shokouhyar S, Tirkolaee EB, Pishva N. Applications and emerging trends of blockchain technology in marketing to develop Industry 5.0 Businesses: A comprehensive survey and network analysis. Internet of Things 2024;28:101401. https://doi.org/10.1016/j.iot.2024.101401.
21. Singh SK, Lee C, Park JH. CoVAC: A P2P smart contract–based intelligent smart city architecture for vaccine manufacturing. Computers & Industrial Engineering 2022;166:107967. https://doi.org/10.1016/j.cie.2022.107967.
22. Toufaily E. An integrative model of trust toward crypto–tokens applications: A customer perspective approach. Digital Business 2022;2:100041. https://doi.org/10.1016/j.digbus.2022.100041.
23. Rajak M, Shaw K. An extension of technology acceptance model for mHealth user adoption. Technology in Society 2021;67:101800. https://doi.org/10.1016/j.techsoc.2021.101800.
24. Mao S, Han X, Lu Y, Wang D, Su A, Lu L, et al. Multi sensor fusion methods for state of charge estimation of smart lithium–ion batteries. Journal of Energy Storage 2023;72:108736. https://doi.org/10.1016/j.est.2023.108736.
25. Guo C, Ke Y, Zhang J. Digital transformation along the supply chain. Pacific–Basin Finance Journal 2023;80:102088. https://doi.org/10.1016/j.pacfin.2023.102088.
26. Dixit VK, Malviya RK, Kumar V, Shankar R. An analysis of the strategies for overcoming digital supply chain implementation barriers. Decision Analytics Journal 2024;10:100389. https://doi.org/10.1016/j.dajour.2023.100389.
27. Thakur P, Kumar Sehgal V. Emerging architecture for heterogeneous smart cyber–physical systems for industry 5.0. Computers & Industrial Engineering 2021;162:107750. https://doi.org/10.1016/j.cie.2021.107750.
28. Marinković M, Al–Tabbaa O, Khan Z, Wu J. Corporate foresight: A systematic literature review and future research trajectories. Journal of Business Research 2022;144:289–311. https://doi.org/10.1016/j.jbusres.2022.01.097.
29. Busboom A. Automated generation of OPC UA information models — A review and outlook. Journal of Industrial Information Integration 2024;39:100602. https://doi.org/10.1016/j.jii.2024.100602.
30. Rathore RK, Mishra D, Mehra PS, Pal O, Hashim AS, Shapi'i A, et al. Real–world model for bitcoin price prediction. Information Processing & Management 2022;59:102968. https://doi.org/10.1016/j.ipm.2022.102968.
31. Noriega R, Pourrahimian Y. A systematic review of artificial intelligence and data–driven approaches in strategic open–pit mine planning. Resources Policy 2022;77:102727. https://doi.org/10.1016/j.resourpol.2022.102727.
32. Wang J, Omar AH, Alotaibi FM, Daradkeh YI, Althubiti SA. Business intelligence ability to enhance organizational performance and performance evaluation capabilities by improving data mining systems for competitive advantage. Information Processing & Management 2022;59:103075. https://doi.org/10.1016/j.ipm.2022.103075.
33. Ran R, Wang X, Wang T, Hua L. The impact of the digital economy on the servitization of industrial structures: the moderating effect of human capital. Data Science and Management 2023;6:174–82. https://doi.org/10.1016/j.dsm.2023.06.003.

34. A. S, Vairavasundaram S, Kotecha K, V. I, Ravi L, Selvachandran G, et al. Blockchain–based trust mechanism for digital twin empowered Industrial Internet of Things. Future Generation Computer Systems 2023;141:16–27. https://doi.org/10.1016/j.future.2022.11.002.
35. Khan M, McNally C. Recent developments on low carbon 3D printing concrete: Revolutionizing construction through innovative technology. Cleaner Materials 2024;12:100251. https://doi.org/10.1016/j.clema.2024.100251.
36. Lu Y, Xiao J, Li Y. 3D printing recycled concrete incorporating plant fibres: A comprehensive review. Construction and Building Materials 2024;425:135951. https://doi.org/10.1016/j.conbuildmat.2024.135951.
37. Wang X, Li W, Guo Y, Kashani A, Wang K, Ferrara L, et al. Concrete 3D printing technology for sustainable construction: A review on raw material, concrete type and performance. Developments in the Built Environment 2024;17:100378. https://doi.org/10.1016/j.dibe.2024.100378.

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.