

Article

Not peer-reviewed version

The Role of Frequent Term Sets in Enhancing Term Co-Occurrence Network Analysis

[Boris Chigarev](#) *

Posted Date: 12 February 2026

doi: 10.20944/preprints202602.0893.v1

Keywords: IEEE term; artificial intelligence; bibliometric analysis; sets of terms; VOSviewer



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.

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.

Article

The Role of Frequent Term Sets in Enhancing Term Co-Occurrence Network Analysis

Boris Chigarev

Institute of Oil and Gas Problems Russian Academy of Sciences; bchigarev@ipng.ru

Abstract

Relevance of this work is determined by the fact that despite the widespread use of keyword sets as the most common approach for collecting thematic information, there are few publications dedicated to the study of frequent term sets in bibliometric research. Usually, pairs of terms co-occurrence are used to construct the network, as in VOSviewer. **Research objective.** 1. Testing the impact of adjusting the construction of the IEEE term co-occurrence graph by increasing the significance of "strong links," which often form sets of multiple terms. 2. Identify IEEE Terms describing a relevant topic more commonly encountered in newer publications. **Materials and methods.** The study used 7,114 bibliometric records from IEEE Xplore for the years 2021-2025, collected based on the query: "IEEE Terms": Artificial Intelligence. Mapping of IEEE terms was performed using VOSviewer, and the FP Growth algorithm was used to identify frequently occurring sets. **Results and conclusions.** Even the simplest enhancement of the significance of terms forming frequently occurring sets showed that the dominant term "artificial intelligence" moved from a cluster with more general words to a cluster with more theme-related terms. An additional result of the research was the identification of a growing interest in the topic described by the terms: *artificial intelligence, training, accuracy, data mining, adaptation models, transformers and vectors*, which seems to be a clear and consistent topic. **Future research.** The author believes that the terms forming frequently occurring sets are important for explaining research topics. Therefore, it is advisable to study the same bibliometric data, but using hypergraphs to represent sets of co-occurring terms.

Keywords: IEEE term; artificial intelligence; bibliometric analysis; sets of terms; VOSviewer

Introduction

Motivation for Conducting this Research

Despite the fact that using keyword sets is the most common approach for collecting thematic information, there are few publications dedicated to examining a set of terms as hyperedges of a hypergraph in bibliometric studies. For example, searching for the word "hypergraph" in the titles and abstracts of the leading scientometrics journal *Scientometrics* yielded only 7 publications (as of January 21, 2024, source: dimensions.ai). At the same time, searching VOSviewer term in the same source yielded 50 publications. Therefore, before directly applying hypergraphs to bibliometric tasks, this paper proposes to first enhance the influence of terms that can form hyperedges (i.e., constitute sets of 3 or more terms) when constructing a term co-occurrence network using VOSviewer. And only after confirming the concept of this approach is it planned to move on to using hypergraphs in bibliometric analysis, conducting a separate study after obtaining minimal confirmation of the proposed concept. Note that when implementing a hypergraph layout, the most common approach is to first construct a bipartite graph, and then the hyperedges are used as attraction centers for the nodes. By increasing the significance of the terms that make up sets of 3 or more terms, we also form additional centers of attraction from them, which in a certain sense is close to some methods of embedding hypergraphs.

To verify that the estimates obtained for the journal "Scientometrics" are not one-sided, let us utilise the fact that the Scilit database has a "Scientometrics & Research Ethics" section, to which the query for "VOSviewer" in the titles, abstracts, and keywords yielded 6303 publications, while a search for "hypergraph" yielded 69 publications.

The choice of a set of key terms is essential for conducting this study. It is desirable that such terms relate to the current research topic (in this case, the AI topic was chosen), originate from a controlled thematic dictionary (expert choice, uniformity of spelling) and are publicly available. These conditions are satisfied by the IEEE Term of the IEEE Xplore reference database.

Despite the high quality of the IEEE Term, they are rarely used in bibliometric analysis. This can be shown by several examples of queries to open reference databases:

- ScienceDirect in Title, abstract, keywords: "IEEE Term" → No results found;
- Dimensiona.ai "IEEE Terms" in Title and abstract → 11 publications of these, 8 links to preprints and articles by the author of this study.
- The request to Scilit "IEEE Terms" did not bring anything new in comparison with the results of Dimensiona.ai.

Thus, it can be argued that "IEEE Terms" are rarely used in bibliometric analysis.

Brief Literature Review

The found publications mentioning the use of hypergraphs in bibliometric analysis argue that hypergraph models offer a more straightforward representation of scientific output and author collaborations than conventional graphs.

Examples of solved problems include:

Citation networks as hypergraphs. Traditional network models typically use graph theory to treat articles as nodes and citations as pairwise relationships between them. In the article [1], an alternative evolutionary model based on hypergraph theory is proposed, in which a single hyperedge (article) can contain an arbitrary number of nodes (citations).

Authors as hypergraph structures. In the article [2], it is proposed to use the hypergraph model — a generalized network — to represent publication data, considering articles as nodes of a hypergraph. The hyper-edges connecting the nodes represent the authors linking all their works. The authors show that this representation is simpler than other models of authorship networks.

Entities as hyperedges or within hyperedges. The hypergraph-of-entity serves as a representation model for terms, entities, and their relationships, functioning as an indexing method in entity-oriented search. A new mixed hypergraph density measure is introduced [3], analyzed through corresponding bipartite mixed graphs. Findings reveal that hyperedge-based node degrees follow a power law distribution, while node-based degrees and hyperedge cardinalities are log-normally distributed.

General hypergraph inference for higher-order relational data. The article [4] presents a framework using statistical inference to analyze the structure of hypergraphs. The method helps identify missing hyperedges of different sizes and discover overlapping communities from complex interactions. It features an efficient numerical implementation, outperforming dyadic algorithms in speed when working with higher-order data.

Several publications reference IEEE terms; however, these terms are typically mentioned without being the focus of bibliometric studies.

Thus, the annotation to [5] simply states that "is the very first issue that is published, speaking in IEEE terms, as an electronic only media publication". In the annotation to [6] there is a line "Today, both the IEEE term "load tap changer (LTC)" and the IEC term "on-load tap-changer (OLTC)" are in the terminology of international standards".

There are also informative articles, the study [7] investigate how authors use keywords, IEEE terms, and social tags in scientific papers from the Faculty of Electrical Engineering and Computing. The results show little connection between these types of keywords, indicating that authors might not see the value of using controlled vocabulary for better information retrieval. This article is very

relevant in the context of our research, which is confirmed by the phrase "the value of using controlled vocabulary for better information retrieval".

In order not to abuse self-citations, let us cite only three of the eight references to my own works that reveal the importance of IEEE terms in bibliometric research.

Article [8] analyzes in detail the use of the IEEE Terms field in IEEE Xplore data to identify publication themes. It is emphasized that this field is an analogue of "Index Keywords" in Scopus and allows high-quality data visualization using VOSviewer and Scimago Graphica.

The study [9] focuses on the limitations and advantages of using IEEE Terms when analyzing 12,000 bibliometric records. The work examines how a controlled vocabulary helps identify trends in energy technologies.

The paper [10] focuses on identifying keywords useful to experts based on data from the IEEE Terms field.

Thus, it can be argued that "IEEE Terms" are used quite rarely when conducting bibliometric analysis.

One aspect that is important for this study is the verification of using additional sets of three or more terms to construct a co-occurrence network using VOSviewer.

The goal is to avoid the classic manifestation of the "dominant context effect" when optimising modularity. In bibliometrics, this happens when a commonly used term (e.g., "Machine Learning") has hundreds of weak links that, in total, outweigh a single strong link to a highly specialized term.

The analysis of network structure typically relies on two primary metrics: betweenness centrality and modularity. The authors [11] explore additional analytical techniques, which include examining average weight concerning endpoint degree, *average weighted nearest neighbor's degree*, weighted clustering coefficient relative to degree, and strength as related to node degree. Here it is worth paying attention to the "average weighted nearest neighbor's degree" meaning the preference is not for all connections, but for local ones.

The Leiden algorithm used in VOSviewer maximizes global modularity. If the term A is strongly related to B , but A has many weak connections with "one cloud of words", and B — with "the second cloud of words", the algorithm can separate them into different clusters. Leyden's logic: The total contribution of many weak connections to the overall density of the graph turns out to be higher than the contribution of one strong connection between A and B .

A common approach in network analysis is the optimization of modularity to divide networks into communities. However, the authors [12] argue that this method may not effectively identify smaller groups, as its effectiveness relies on the overall network size and the interconnection of modules, even if those modules are well-defined.

Research [13] shows that five algorithms—Leiden (Constant Potts Model), Leiden (modularity), Infomap, Markov Cluster, and Iterative k-core—often find communities that lack proper connections. To address this, a Connectivity Modifier (CM) has been developed to remove small edge cuts and re-cluster until the communities meet specific well-connectedness standards.

Leiden focuses on improving the modularity function for graph clustering, but its greedy method can quickly lead to less effective solutions. To address this, researchers [14] have developed a new method, called Tel-Aviv University (TAU), which uses a genetic algorithm to better explore the solutions. TAU is designed to optimize modularity and can avoid getting stuck in local optima, *making it especially useful when there are many connections between communities*, as this often complicates community partitioning in complex systems. The algorithm emphasizes Jaccard similarity of over 0.98 between the best groupings from different generations. Explanation of the above: "many connections between communities... in complex systems" is a fairly common phenomenon when clustering keywords using VOSviewer, related to the presence of terms that have many weak connections. For example, the term "energy" can appear in combination with terms from various thematic clusters. The Jaccard similarity criterion can reflect the relatedness of several terms simultaneously, strengthening the significance of multiple co-occurrences.

More indirectly, some issues with using the Leiden clustering method are reflected in the article [15], which evaluates the clusters from the Leiden and Iterative K-core clustering algorithms to see how easily they can become disconnected by removing a few edges. It notes that most clusters from Leiden clustering in real-world networks, except when using high resolution parameters, do not fit the criteria for well-connected clusters. Additionally, it finds that the final clustering on real-world networks include only a small number of nodes, indicating that not all nodes belong to communities.

One approach to solving the problem is to adjust the data so that if a pair or more terms form a "hard" micro-cluster, the algorithm is more likely to keep them together. To achieve this, our work employs an approach to increase the significance of strongly connected terms at the level of their inclusion in the graph.

Research Objective

1. Test the significance of adjusting the IEEE Terms co-occurrence graph construction by adjusting the original data by two methods: increasing the significance of "strong links" by adding "virtual records" containing strongly connected terms and filtering the original term list, leaving those that often form sets of several terms.
2. Identify IEEE Terms describing a relevant topic more commonly encountered in newer publications.

Materials and Methods

The materials for the analysis were IEEE Xplore bibliometric data for 2021-2025, collected on request: "IEEE Terms": Artificial Intelligence. Filter used: journal articles only. A total of 7,114 records were exported. Current as of Jan 19, 2026.

We applied the FP Growth algorithm [16] to lists of IEEE terms to identify frequent co-occurring sets of 2 to 5 terms.

IEEE Terms mapping was performed by VOSviewer [17] a freely distributed program for constructing and visualizing bibliometric networks.

Results and Discussions

Key Characteristics of the Records Used

The bibliometric data exported from IEEE Xplore is of very high quality, only 5 of the 7114 records had no DOI.

Despite specific time filtering, some records did not fall within the 2021-2025 timeframe. Thus, 359 entries in the "Publication Year" field related to the year 2026.

The exact distribution by year in the exported records was: 2021 (417); 2022 (657); 2023 (867); 2024 (1410); 2025 (3361), the rest belonged to other years.

It is noticeable that in 2025 the number of publications on the topic under study increased sharply.

We are interested in the topic of Artificial Intelligence as defined by the "IEEE Terms," rather than the distribution by year, so all records were used in the study. All entries in the "IEEE Terms" field were filled in.

If considering the cases of complete coincidence of the entries in the "IEEE Terms" field, then 28 times it had the value "Training; Data mining; Artificial intelligence", 7 times — "Artificial intelligence", 5 — "Sun; Wireless communication; Wide area networks; Information science; Artificial intelligence" — specifically, entries with this set of terms referred to the "Editorial board" of the "Journal of Communications and Networks". Other sets of "IEEE Terms" were mostly encountered only once. Considering that there were 7114 records in total, the number of completely matching records was not large.

Most of the entries with the set "IEEE Terms": "Training; Data mining; Artificial intelligence" belonged to the journal *IEEE Transactions on Instrumentation and Measurement*. This suggests that a separate study of the distribution of "IEEE Terms" across journals might make sense.

On average, there were 9 "IEEE Terms" in the entries, which exceeds the average number of author keywords — 5.3.

It was precisely the high average number of "IEEE Terms" that made it possible to search for sets of co-occurring 2–5 terms in the records.

For comparison, the field of author keywords was filled in 6768 entries out of 7114.

The main sources of publications were: IEEE Access (1958), IEEE Transactions on Artificial Intelligence (380), IEEE Internet of Things Journal (359), IEEE Transactions on Engineering Management (180), IEEE Transactions on Instrumentation and Measurement (149), IEEE Transactions on Consumer Electronics (146), IEEE Transactions on Intelligent Transportation Systems (117), IEEE Sensors Journal (108) and IEEE Transactions on Geoscience and Remote Sensing (100). The number of publications is given in parentheses.

Building IEEE Terms Co-occurrence Networks Using VOSviewer

Only IEEE Terms were used in this study. The choice is due to the fact that these terms are included in the controlled vocabulary reviewed and updated by IEEE experts. It is accessible for use, and the spelling of the terms is formalized, with rare exceptions. That is, such terms can be applied without prior preparation, such as lemmatization. This allows us to focus on the problem of the influence of additional modification of the list of terms involved in constructing their co-occurrence network. The list is modified, but not the terms themselves. Another important factor is the ability to search in IEEE Xplore using IEEE Terms as part of the metadata of bibliometric records.

Three methods were compared in order to determine the significance of the change to the list of keywords used, as described sequentially below.

1. The usual approach to using VOSviewer in bibliometric analysis. The main parameters are applied as default, except that in this case, only 302 keywords were used instead of the standard 1000. This is because the number of terms used in constructing the co-occurrence network of keywords must be the same in all three cases. This is due to the requirement that the co-occurrence network of keywords in all three scenarios should have the same number of terms.
2. The option where several "virtual records" containing IEEE Terms forming groups of 4 or 5 terms are added to the bibliometric records.
3. The option where only the IEEE Terms are left, forming a group of two or more terms with 0.5% support.

The IEEE Terms field was used as the Index Keywords field in Scopus data. The visualization results of the first network option are shown in Figure 1. Next, we will call this option Default.

Each of the three figures consists of two parts: the upper one reflects the clustering of IEEE Terms based on their co-occurrence, and the lower one is the overlay Avg. pub. year, showing the occurrence of terms over time.

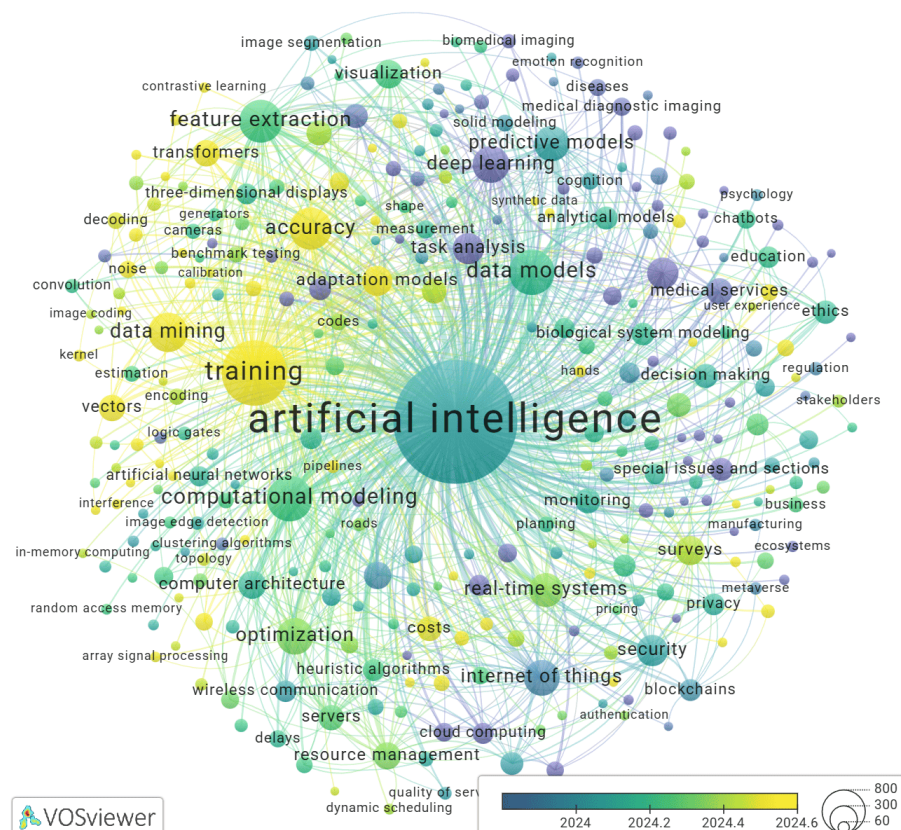


Figure 1. The 302 IEEE Terms co-occurrence network with the highest Total link strength value built with other default settings. The list of IEEE Terms used has not changed. Items: 302 | Links: 22291 | Total link strength: 136202 | Clusters: 6.

The primary term used in this study is *artificial intelligence*, so here is a list of the most frequently occurring terms that form a cluster with it (red). artificial intelligence (37668), machine learning (2613), task analysis (2214), surveys (2473), costs (1757), decision making (1687), reviews (1622), ethics (1590), analytical models (1544), special issues and sections (1079), biological system modeling (1495), collaboration (1421), technological innovation (1382), education (1262), games (1092), robots (1042), safety (1073), uncertainty (1063), market research (1051). It is noteworthy that these terms are quite general in nature.

At the bottom of the figure, you can see the terms most frequently found in new publications (Avg. pub. year). Here is a list of them and their characteristics, as you can get by uploading the clustering data obtained in VOSviewer to the online service <https://app.vosviewer.com/>. All terms belong to the green cluster, which does not include the term *artificial intelligence*.

- Item: artificial intelligence | Links: 301 | Total link strength: 37668 | Occurrences: 6984 | Avg. pub. year: 2024.08
- Item: training | Links: 301 | Total link strength: 11918 | Occurrences: 1950 | Avg. pub. year: 2024.68
- Item: data mining | Links: 281 | Total link strength: 4006 | Occurrences: 707 | Avg. pub. year: 2024.86
- Item: accuracy | Links: 293 | Total link strength: 5313 | Occurrences: 807 | Avg. pub. year: 2024.86
- Item: transformers | Links: 231 | Total link strength: 2129 | Occurrences: 319 | Avg. pub. year: 2024.75
- Item: vectors | Links: 244 | Total link strength: 1836 | Occurrences: 316 | Avg. pub. year: 2024.97
- Item: adaptation models | Links: 278 | Total link strength: 2821 | Occurrences: 413 | Avg. pub. year: 2024.66

The simplest way to increase the significance of terms that are grouped together is to add "virtual entries" to the list of existing bibliometric records, which will contain exactly those terms in the IEEE Terms field. 40 "virtual entries" were added, containing the most frequently forming groups of 4 or 5 terms. The search for combinations of such terms is implemented using the FP Growth utility.

Otherwise, in this case, the construction of the IEEE terms co-occurrence network was performed similarly to the previous option. The results are shown in Figure 2.

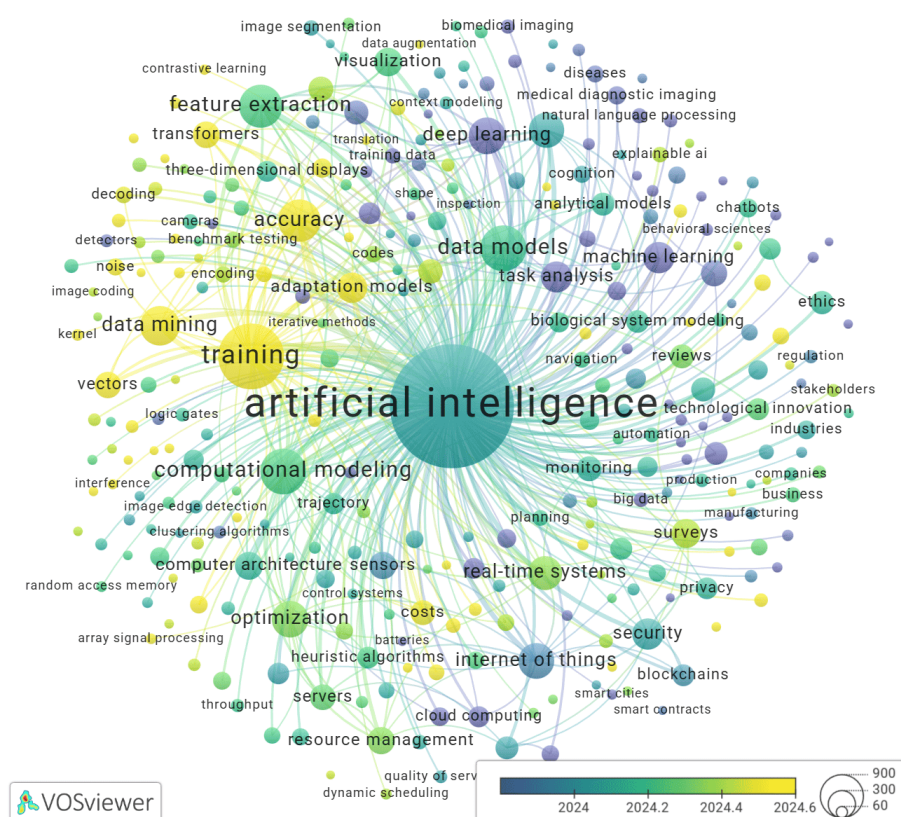


Figure 2. The 302 IEEE Terms co-occurrence network, built by adding 40 "virtual records" containing groups of 4 and 5 co-occurring terms. Items: 302 | Links: 22291 | Total link strength: 136446 | Clusters: 4.

The first thing that is obvious is that when adding such "virtual records", the number of clusters decreased from 6 to 4.

At the same time, the main term (artificial intelligence) used in this work has moved to another cluster, many of which terms refer to the green cluster shown in Figure 1 and the red one in Figure 2. Note: By default, the colors for clusters are in the following sequence: red, green, blue, khaki.

The following are the characteristics of the term *artificial intelligence* and the 6 terms most frequently used in new publications.

Item: artificial intelligence | Links: 301 | Total link strength: 37786 | Occurrences: 7023 | Avg. pub. year: 2024.09
Item: training | Links: 301 | Total link strength: 12003 | Occurrences: 1978 | Avg. pub. year: 2024.68
Item: data mining | Links: 281 | Total link strength: 4027 | Occurrences: 714 | Avg. pub. year: 2024.86
Item: accuracy | Links: 293 | Total link strength: 5343 | Occurrences: 817 | Avg. pub. year: 2024.87
Item: transformers | Links: 231 | Total link strength: 2141 | Occurrences: 323 | Avg. pub. year: 2024.75
Item: vectors | Links: 244 | Total link strength: 1839 | Occurrences: 317 | Avg. pub. year: 2024.97
Item: adaptation models | Links: 278 | Total link strength: 2836 | Occurrences: 418 | Avg. pub. year: 2024.67

For the terms *artificial intelligence*, *accuracy*, and *adaptation models*, the "Avg. pub. year" parameter is slightly higher compared to the previous version. A likely explanation is that the "virtual records" referred to a more recent time period, but given that only 40 entries were added in 7114, the changes are minor.

This means that the arbitrary formation of "virtual record" fields that are not related to IEEE Terms can affect the calculated parameters. Therefore, it is rational to consider a modification that increases the significance of local joint terms building a network of terms without including "virtual entries." But here a problem arises. In all 7114 IEEE Terms records contain 3619 unique terms, if we consider only those that meet the 0.5% support for groups of terms from three to five, then there will be only 95 unique terms. That is, the joint occurrence of 3 or more terms in a group, even with 0.5% support, severely limits the number of unique terms. To expand the list of unique terms for building a network, we will find groups of terms starting from 2. With 0.5% support, there were only 302 such unique terms. We will use them when building a network in the next step of the study.

Note: There are many options for narrowing the sample can be proposed, for example, reducing the level of support, but increasing the minimum joint occurrence to 3. But this paper did not aim to conduct a detailed analysis of possible options. It was necessary to demonstrate the significance of local, "strong" links for networking and term clustering. The limit of 302 terms in the network in previous versions was chosen for comparability with this version of the network.

Figure 3 shows the term co-occurrence network when only 302 unique terms remain in the "IEEE Terms" field.

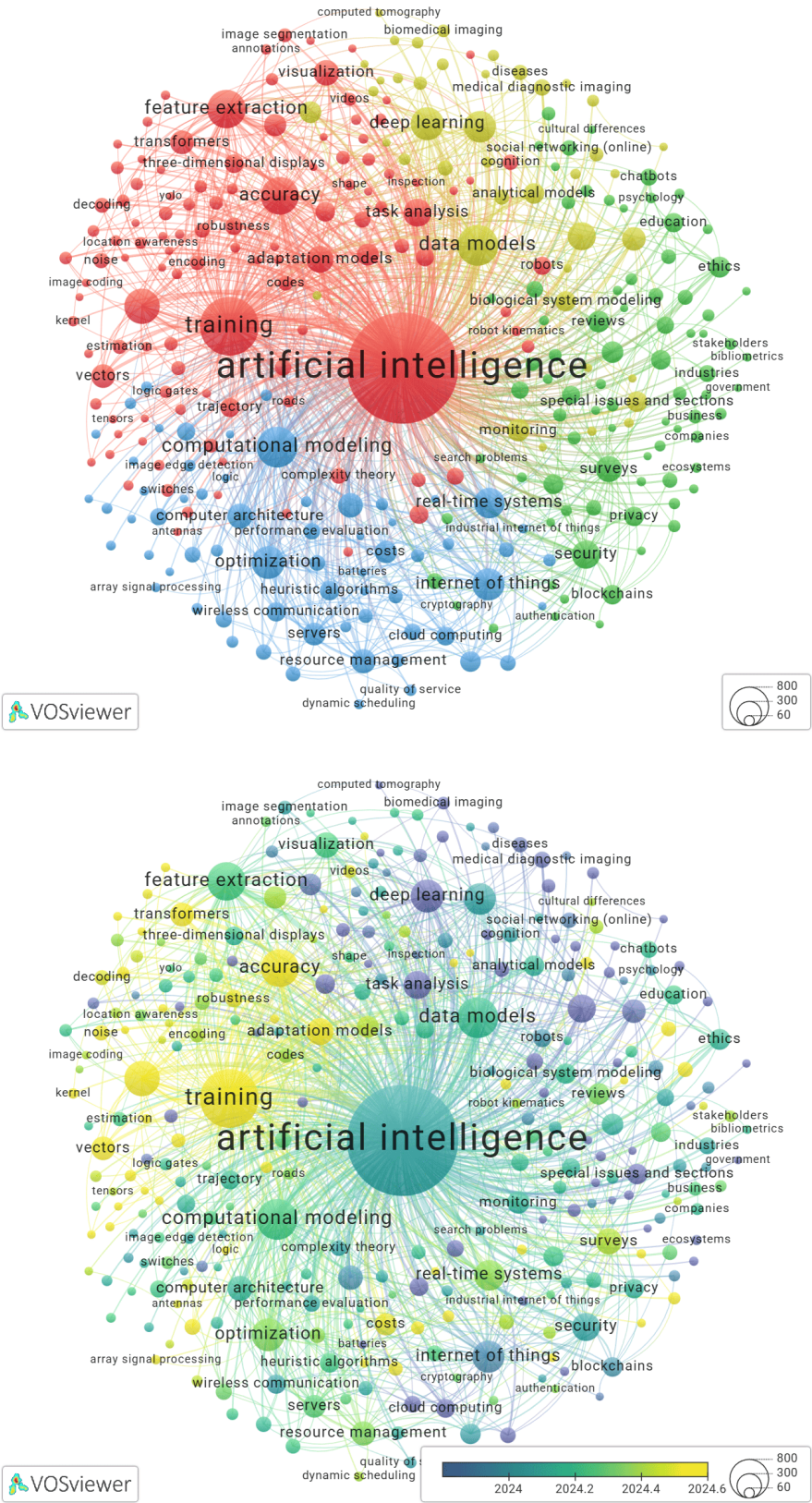


Figure 3. IEEE Terms co-occurrence network **constructed from 302 terms extracted from** frequent itemsets with 0.5% support. Items: 302 | Links: 22156 | Total link strength: 135899 | Clusters: 4.

Let us take a closer look at how the data was obtained to construct this variant of the IEEE Terms co-occurrence network.

All 7114 entries of the IEEE Terms field are taken. Using FP Growth, a list of term sets consisting of 2-5 terms was obtained with 0.5% support. That is, the parameter -s0.5m2n5 was used for the fpgrowth utility. A list of terms included in the resulting output file has been compiled from it. Next, the terms in the list were deduplicated. There were 302 unique IEEE Terms in total. To build a network containing only these 302 terms, the remaining terms were excluded from consideration using a file thesaurus_terms.txt . It includes all the unique IEEE Terms from 7114 entries, with the exception of 302.

As with the "virtual entries" option, the term *artificial intelligence* entered the first (red) cluster. The following are the characteristics of the term *artificial intelligence* and the 6 terms most frequently used in new publications.

Item: artificial intelligence | Links: 301 | Total link strength: 37690 | Occurrences: 6984 | Avg. pub. year: 2024.08
Item: training | Links: 301 | Total link strength: 11911 | Occurrences: 1950 | Avg. pub. year: 2024.68
Item: data mining | Links: 282 | Total link strength: 4003 | Occurrences: 707 | Avg. pub. year: 2024.86
Item: accuracy | Links: 294 | Total link strength: 5326 | Occurrences: 807 | Avg. pub. year: 2024.86
Item: transformers | Links: 231 | Total link strength: 2131 | Occurrences: 319 | Avg. pub. year: 2024.75
Item: vectors | Links: 246 | Total link strength: 1837 | Occurrences: 316 | Avg. pub. year: 2024.97
Item: adaptation models | Links: 278 | Total link strength: 2824 | Occurrences: 413 | Avg. pub. year: 2024.66

Unlike the second option, there are no changes in the Avg. pub. year parameter compared to the default. Therefore, this option is preferable to the second one.

Compared to the default option, the first 6 terms in this list have a slightly larger Links parameter, while for *adaptation models* this parameter remains the same. The difference in the *Total Link Strength* parameter can be in either direction.

Discussion of the Results

The qualitative difference between the network of co-occurrence of terms in the original data and the data with enhancement of the significance of the terms forming frequently occurring sets is that in the first case, the dominant term *artificial intelligence* refers to a cluster containing the terms: *machine learning, task analysis, surveys, decision making, reviews, ethics, analytical models*, i.e., a more general traditional topic related to AI. In the third case, *artificial intelligence* belongs to a cluster that includes terms such as *training, feature extraction, accuracy, data mining, task analysis, adaptation models, and visualization* — that is, a more specific topic. Moreover, as the Overlay maps show, these terms are more common in newer publications.

If comparing options 1 and 2, then the second variant adds "virtual records" in which the terms forming frequently occurring sets are used. That is, the number of unique terms remains the same, but the total number of links has increased. This corresponds to the fact that *Links: 22291* is the same in both cases, and the *Total link strength* in the first case is 136202 vs 136446 in the second case. This might explain why, due to a slight increase in the overall link strength, 6 clusters were obtained in the first case and 4 in the second.

In the third case, preference is given to terms that form frequently occurring sets. With the same number of terms (302), this led to a decrease in the number of unique paired links between terms to 22156, and the total number of links (Total link strength) 135899 became even lower than in the first case. The number of clusters in this case became equal to 4, as in the second case. This shows that the attempt to give preference to the terms forming frequently occurring sets was successful. The latter is important for describing topics. Two terms poorly reflect the specifics of the topic, while 3–5 terms provide a significantly more accurate description.

To assess the significance of terms describing a newer topic (Overlay graphs with Avg. pub. year), we will consider the parameters of the five most frequently encountered terms, which will be discussed in order of frequency of occurrence (Occurrences value).

Item: artificial intelligence. This term is linked to all other terms in the column in all cases (Links: 301). In the second variant, this term occurs more frequently than in the other two (Occurrences: 6984[1], 7023[2], 6984[3]), which is understandable, since "virtual records" containing multiple co-occurring terms were added. Accordingly, the Total link strength for the second case became the highest (37668[1], 37786[2], 37690[3]). A slight increase in the third variant compared to the first indicates that the significance of the term artificial intelligence for this variant is higher.

It is worth noting that the use of "virtual records" in the second variant affected Avg. pub. year. This distortion makes the second option less preferable for analysis.

Item: training. Like *artificial intelligence*, the term *training* is linked to all other terms, Links (301). The occurrences in the first and third cases are the same (1950[1], 1978[2], 1950[3]), and Total link strength (11918[1], 12003[2], 11911[3]) in the third case, it is lower than in the first. Given the more significant difference between the second option and the first and third, and the fact that in the second and third cases we obtained 4 clusters each, in the future we will pay more attention to comparing the first (let's call it the original) and the third option (without "virtual additions").

Item: accuracy. In this case, "virtual records" also increased the Occurrences parameter for the second case (807[1], 817[2], 807[3]). The most interesting aspect is that the Links parameter (293[1], 293[2], 294[3]) and Total link strength (5313[1], 5343[2], 5326[3]) for the third option are somewhat higher than for the first.

Item: data mining. Here: Occurrences are respectively equal to: (707[1], 714[2], 707[3]), Links — (281[1], 281[2], 282[3]), Total link strength — (4006[1], 4027[2], 4003[3]). Unlike *accuracy*, here in the third case, *Total link strength* is somewhat lower than in the first case.

Item: adaptation models. For this term, the difference between the first and third options is reflected in the *Total link strength* (2821[1], 2836[2], 2824[3]). That is, the variation in *Total link strength* for different terms can have different directions for the first and third cases.

Item: transformers. Here, as in the previous case, the *Total link strength* in the third case is somewhat higher than in the first (2129[1], 2141[2], 2131[3]).

Item: vectors. This term is characterized by the fact that *Links* (244[1], 244[2], 246[3]) for the third case increased more significantly compared to the first, than *Total link strength* (1836[1], 1839[2], 1837[3]).

If we describe the current topic using all the terms at once: artificial intelligence, training, accuracy, data mining, adaptation models, transformers, and vectors, it is unrealistic to find a publication that satisfies all seven terms at once. Therefore, it is advisable to use search engines that utilize AI, such as Semantic Scholar or Elicit. Such search engines will look for publications not by exact term matching, but publications that are close by context.

Below are a few examples. To demonstrate the relevance of the publications founded, we will only show short lines from the abstracts containing the terms founded in them.

Thus, Semantic Scholar suggested the following publications that are most relevant according to its criteria. In the abstract of the first one [18], there are lines: "model driven by artificial intelligence", "optimize data collection and processing using artificial intelligence technology", "social adaptation and contribution", "talent training model", "classification accuracy", "processing learning behavior data". In the second article [19], the abstract contained the following lines: "Artificial Intelligence and automation", "data mining plays a crucial role", "Transformers (BERT) model", "Career Success data set", "machine learning models such as support vector machines", "classification accuracy".

Example proposed by Elicit. In the annotation [20], there are lines: "Artificial Intelligence such as NeurIPS and ICLR", "machine learning models", "Large Language Mode", "pre-trained word vectors using Word2Vec, GloVe, FastText", "transformer architecture-based BERT, DistilBERT", "TF-IDF vectorize", "Support Vector Machines". This differs from the seven terms under consideration, but if we go beyond the annotation, the full text contains: "massive pre-training," "training dataset," "test accuracy." The term "adaptation models" is not directly found in the text, but semantically related terms such as "nearest neighbors technique" and "domain adaptation techniques" are present.

What is common in these publications is that they do not directly address IEEE topics, but this only underscores the importance of an interdisciplinary approach in solving problems related to AI. If earlier, when conducting research intelligence, the focus was on identifying closely related thematic solutions, then in the context of AI, large-scale solutions that can be applied to problems in various fields of knowledge are beginning to play a significant role. High-quality LLMs and the fine-tuning method can be applied to geology, sociology, and energy. In [21] a system using multiple agents to produce high-quality training data for environmental language models was created. Authors developed ChatEnv, which includes a balanced instruction set of 100 million tokens **across five** environmental **themes**. Another example could be methods for vectorizing terms in text; an effective method for vectorizing terms can be extended to other categorical data by using their characteristic context. In the work [22], the authors adapted the Word2Vec architecture to work with categorical features in tabular data.

The problem with the Leiden/Louvain algorithms in the context of bibliometrics and text analysis is that modularity is a "global" measure. Its goal is to group nodes in such a way that the density of connections within a cluster is higher than would be expected by chance, which may lead to sacrificing strong local connections (specific pairs of terms) for the sake of the overall graph structure. On the other hand, strong links between a pair (or more) of terms will be important for subsequent information gathering on the topic of interest, which means they are indeed significant for analyzing current themes. Unlike Leiden, KaHIP-type tools [23] are focused on minimizing cuts (Min-Cut). This preserves strong pairs: if two terms have a very high connection weight (they often appear together), KaHIP is likely to leave them in the same cluster, as breaking such a link is "expensive" for its objective function. However, in Leiden, such terms can diverge if each of them has even stronger (collectively) connections to other "clouds of terms." KaHIP will create clusters of approximately equal size.

When using VOSviewer, a large topic containing general terms and several smaller clusters may emerge. KaHIP breaks down a large topic into smaller, comparable subdomains. This makes the description more detailed, but sometimes it artificially divides entire areas. In our work, by focusing on sets of terms that often occur together, we artificially enhance the role of local connections in terms of assessing overall links.

Considering the above, it is advisable to conduct a separate study to compare the result obtained in this study and the analysis, in which frequent sets of terms will be considered as hyperedges.

Conclusions

Even the simplest enhancement of the significance of the terms forming frequently occurring sets showed that the dominant term "artificial intelligence" moved from a cluster with more general words to a cluster with more thematically related terms.

The proposed approaches to increasing the significance of terms forming frequently occurring sets, with a slight change in the main network indicators: *Occurrences*, *Links*, and *Total link strength*, lead to a significant change in clustering results, reducing the number of clusters in the network from 6 to 4.

The option with the preference of terms forming frequently occurring sets gave a similar result to the option of adding "virtual records", but in this case did not lead to a distortion of such a calculated parameter as *Avg. pub. year*. Therefore, this method will be preferable for use in subsequent studies.

This article does not aim to verify the proposed approach to data modification for constructing term co-occurrence networks on various samples. The work was conducted as a proof of concept to explore adjustments to a known method for constructing term co-occurrence networks by enhancing the significance of terms forming frequently occurring sets, as it can improve the description of current research topics.

An additional result of the study was the identification of an increasing interest in the topic described by the terms: *artificial intelligence, training, accuracy, data mining, adaptation models, transformers and vectors*, which seems to be a clear and consistent topic.

Given the significance of terms forming frequently occurring sets for describing research topics, the author considers it appropriate to conduct a separate study on the same sample of bibliometric data, but using hypergraphs in which hyperedges represent sets of co-occurring terms.

Acknowledgements. This work was funded by the Ministry of Science and Higher Education of the Russian Federation, State Assignment no. 125021302095-2

References

1. [1] Hu F, Ma L, Zhan X-X, Zhou Y, Liu C, Zhao H, Zhang Z-K. The aging effect in evolving scientific citation networks. *Scientometrics* 2021;126:4297–309. <https://doi.org/10.1007/s11192-021-03929-8>.
2. [2] Lung RI, Gaskó N, Suci MA. A hypergraph model for representing scientific output. *Scientometrics* 2018;117:1361–79. <https://doi.org/10.1007/s11192-018-2908-2>.
3. [3] Devezas J, Nunes S. Characterizing the hypergraph-of-entity and the structural impact of its extensions. *Appl Netw Sci* 2020;5:79. <https://doi.org/10.1007/s41109-020-00320-z>.
4. [4] Contisciani M, Battiston F, De Bacco C. Inference of hyperedges and overlapping communities in hypergraphs. *Nat Commun* 2022;13:7229. <https://doi.org/10.1038/s41467-022-34714-7>.
5. [5] Vlacic L. On Intelligent Transportation Systems Vulnerability [Editor's Column]. *IEEE Intell Transport Syst Mag* 2023;15:4–5. <https://doi.org/10.1109/MITS.2022.3220827>.
6. [6] Harlow JH, editor. Load Tap Changers. *Electric Power Transformer Engineering*. 0 ed., CRC Press; 2007, p. 311–42. <https://doi.org/10.1201/9781420008715-17>.
7. [7] Vrkic D. Are they a perfect match? Analysis of usage of author suggested keywords, IEEE terms and social tags. 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia: IEEE; 2014, p. 732–7. <https://doi.org/10.1109/MIPRO.2014.6859662>.
8. [8] Chigarev BN. On Visual Data Analysis of IEEE Xplore Bibliometric Records on Machine Learning and Artificial Intelligence for Power Systems. *Energy Systems Research* 2025;8:12–29. <https://doi.org/10.25729/esr.2025.01.0002>.
9. [9] Chigarev B. Analysis of the Use of Author Keywords and IEEE Terms in IEEE Xplore Data to Identify Current Research Topics in Energy Technology and Existing Limitations 2025. <https://doi.org/10.20944/preprints202510.0086.v1>.
10. [10] Chigarev BN. IEEE Terms Analysis of 2019-2024 IEEE Xplore Data on the Topic of Energy Systems. *Energy Systems Research* 2024;7:26–38. <https://doi.org/10.25729/esr.2024.02.0003>.
11. [11] Radhakrishnan S, Erbis S, Isaacs JA, Kamarthi S. Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PLoS ONE* 2017;12:e0172778. <https://doi.org/10.1371/journal.pone.0172778>.
12. [12] Fortunato S, Barthélemy M. Resolution limit in community detection. *Proc Natl Acad Sci USA* 2007;104:36–41. <https://doi.org/10.1073/pnas.0605965104>.
13. [13] Park M, Tabatabaee Y, Ramavarapu V, Liu B, Pailodi VK, Ramachandran R, Korobskiy D, Ayres F, Chacko G, Warnow T. Well-connectedness and community detection. *PLOS Complex Syst* 2024;1:e0000009. <https://doi.org/10.1371/journal.pcsy.0000009>.
14. [14] Gilad G, Sharan R. From Leiden to Tel-Aviv University (TAU): exploring clustering solutions via a genetic algorithm. *PNAS Nexus* 2023;2:pgad180. <https://doi.org/10.1093/pnasnexus/pgad180>.
15. [15] Park M, Tabatabaee Y, Ramavarapu V, Liu B, Pailodi VK, Ramachandran R, Korobskiy D, Ayres F, Chacko G, Warnow T. Well-Connected Communities in Real-World and Synthetic Networks 2023. <https://doi.org/10.48550/ARXIV.2303.02813>.
16. [16] Borgelt C. An implementation of the FP-growth algorithm. *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, Chicago Illinois: ACM; 2005, p. 1–5. <https://doi.org/10.1145/1133905.1133907>.

17. [17] 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>.
18. [18] Feng Y. Application Research of Dataset Analysis and Optimization Method Based on Artificial Intelligence Technology in University Talent Training System. *Int J Hi Spe Ele Syst* 2025;2540703. <https://doi.org/10.1142/S012915642540703X>.
19. [19] Zihan Z. A multi-factor data mining and transformer-based predictive modeling approach for career success using educational and behavioral traits. *Sci Rep* 2025;15:39484. <https://doi.org/10.1038/s41598-025-23078-9>.
20. [20] Pendyala VS, Kamdar K, Mulchandani K. Automated Research Review Support Using Machine Learning, Large Language Models, and Natural Language Processing. *Electronics* 2025;14:256. <https://doi.org/10.3390/electronics14020256>.
21. [21] Zhang Y, Lin S, Xiong Y, Li N, Zhong L, Ding L, Hu Q. Fine-tuning large language models for interdisciplinary environmental challenges. *Environmental Science and Ecotechnology* 2025;27:100608. <https://doi.org/10.1016/j.es.2025.100608>.
22. [22] Waldemar H, Martin S, Markus W. Word2Vec Embeddings for Categorical Values in Synthetic Tabular Generation. 2022 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA: IEEE; 2022, p. 613–22. <https://doi.org/10.1109/CSCI58124.2022.00114>.
23. [23] Kobourov S, Meyerhenke H, editors. 2019 Proceedings of the Twenty-First Workshop on Algorithm Engineering and Experiments (ALENEX). Philadelphia, PA: Society for Industrial and Applied Mathematics; 2019. <https://doi.org/10.1137/1.9781611975499>.

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.