Article

Keyword Co-Occurrence Analysis Using the FPGrowth Algorithm. An Example of Energies Journal Bibliometric Data for 2023-2024

Boris Chigarev *

Posted Date: 20 June 2024

doi: 10.20944/preprints202406.1380.v1

Keywords: FP-growth algorithm; keyword co-occurrence; bibliometric data; clustering; visualization

 doi:10.20944/preprints202406.1380.v1

Article

Keyword Co-Occurrence Analysis Using the FP-Growth Algorithm. An Example of Energies Journal Bibliometric Data for 2023-2024

Boris Chigarev

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

Abstract: *Background.* Keyword co-occurrence analysis is a crucial tool for comprehending research trends, identifying relevant studies, and gaining insight into the connections between various concepts and topics. *Objective.* This study focuses on analyzing the co-occurrence of keywords using FP-growth algorithm and direct search methods. *Materials and methods*. The methodology involved extracting bibliometric data of Energies journal for 2023-2024 from MDPI publisher platform, keyword lemmatization and keyword co-occurrence estimation. Clustering and visualization were performed using Multidendrograms and Scimago Graphica software. *Results.* The results showed that the FP-growth algorithm can achieve a close match with the direct search results, which facilitates data preparation for clustering. In addition, finding the co-occurrence of three or more keywords significantly reduced the number of possible combinations, which allowed the identification of specific research topics. *Conclusions.* This study highlights the usefulness of the FP-growth algorithm in keyword analysis and provides insights into ways to refine search queries to abstract databases for the purpose of designing and writing literature and systematic reviews.

Keywords: FP-growth algorithm; keyword co-occurrence; bibliometric data; clustering; visualization

Introduction

FP-growth (Frequent Pattern Growth Algorithm) algorithm is a widely used method for finding frequent patterns in data, in which a frequent pattern tree (FP tree) containing frequency and association information is constructed from the input dataset. [1]

An implementation of the algorithm developed by Christian Borgelt [2] was used in this paper. In this paper, the author describes a C implementation of the FP-growth algorithm, in which projected FP-trees are pruned by removing elements that have become sparse due to projection.

To illustrate the scope of use of this algorithm, consider the results of the query "FP-growth" to the ScienceDirect abstract database, for which 982 results are obtained, including Review articles (858) and Research articles (84) respectively. The relevance of this algorithm can be judged by the number of publications by year: 2023 – 75, 2022 – 104, 2021 – 84, 2020 – 85, 2019 – 73 publications.

The query: "FP-growth" AND (keywords OR "key words") AND (co-occurrence OR cooccurrence)" (up to date as of 28.05.2024) yielded 72 results, thus confirming the rationality of using the FP-growth algorithm in keyword analysis.

Of the found publications, 22 are related to the analysis of Twitter short texts, 7 to the analysis of Scopus data and 9 to Web of Science. In most of the publications the FP-growth algorithm is mentioned in the introduction or in the references. Therefore, in addition, the search for publications related to applications and algorithms analyzing keywords of bibliometric data was conducted.

Query on fields Title, Abstract, Keywords: "FP-growth" AND co-occurrence returned 3 results. Among them, the publication [3], which is closest to our work, uses geophysical data from various sources, including meteorological stations and Internet information, to identify unknown relationships between parameters using associative rules.

 doi:10.20944/preprints202406.1380.v1

Given that the Louvain algorithm is most commonly used in text and keyword analysis, the query on fields Title, abstract, keywords "FP-growth" AND "Louvain'' was formulated for which one publication was obtained [4]. In the paper, Louvain was used to calculate the stability and strength of a person's socioeconomic ties. And FP-Growth, used to make a decision about whether to include a person in an inclusive program.

Having in mind that such limited results were obtained that are only indirectly relevant to the problem under consideration, it was decided to look at the results of the broader query to "Title, Abstract, Keywords": "FP-growth" AND "text mining" for which one result was found [5]. In it, the authors tried to parallelize the FP-growth algorithm on multi-core computers, which is useful for detecting frequently occurring keywords in texts.

The above publication turned out to be the closest to the objective we set ourselves. But it primarily concerns the parallelization of the FP-growth algorithm itself, but not its applicability to the analysis of keyword co-occurrence. This task is valid, there are 21 results for the query 'parallel FP-Growth' on Github.com, but in our case the FP-growth implementation given by Christian Borgelt worked very fast and it was not reasonable to switch to parallel FP-Growth[1].

Query to Scilit abstract database: 'Common Fields [Title, Abstract, Keyword] FP-growth AND Common Fields [Title, Abstract, Keyword] "text mining"' allowed to find 19 publications of which [6] is the most interesting for our study. The authors introduce a two-step keyphrase extraction method that uses the FP-growth algorithm to retrieve frequently used neighbor words, followed by meaningful keyphrase extraction using Latent Dirichlet Allocation (LDA) topic modeling.

In our work, we use the FP-growth algorithm to obtain the co-occurrence of keywords but not neighboring words, then we use search (e.g., grep) to refine the co-occurrence of keywords, and then use clustering to identify topics. No studies have been found that address such a task.

Materials and Methods

Data

In this paper, the source of keywords was Energies journal bibliometric data exported from the MDPI publisher platform for the years 2023 and 2024.

Of these for 2023, 7,617 records for Article Type = Research Article, Review Article.

For 2024, 2,142 records, up to date as of May 18, 2024.

All 9759 records had the 'KEYWORDS' field filled in. This is an advantage of publisher platforms as opposed to some refractive open access databases where some fields may be poorly populated (see The Lens, for example).

Reason for choosing Energies journal: matches the scientific interests of the author of this article, has a high ranking, a large number of annual publications and open access to full texts.

Some formal indicators of this journal according to scimagojr.com: H-INDEX = 152, CiteScore category rank: Q1: Control and Optimization; Q1: Engineering (miscellaneous); Q2: Electrical and Electronic Engineering; Q2: Energy Engineering and Power Technology; Q2: Energy (miscellaneous); Q2: Fuel Technology; Q2: Renewable Energy, Sustainability and The Environment.

Here are some characteristics of the 'KEYWORDS' field records for 9759 records:

- Total number of keywords without de-duplication → 52553
- 30678 after a simple deduplication
- 26061 contain spaces, i.e. they are multi-word terms
- 1026 keywords contain abbreviations

Programs and Utilities in Use

Keyword lemmatization was done using a dictionary lemmatizer with 337751 entries, collected on Github and then edited. The advantage of a dictionary-based lemmatizer is that it is easy to edit

[1] https://github.com/search?q=parallel%20FP-Growth&type=repositories

 doi:10.20944/preprints202406.1380.v1

when errors are found or new terms are added. For example, the term TES (thermal energy storage) can be shortened to TE by a rule-based lemmatizer and then simply deleted during text preparation by applying the rule that terms must contain at least three letters. Some lemmatizers persistently abbreviate biogas to bioga, which is not so critical when doing analytics, but bad when visualizing results. The use of rules is simply necessary in large, regular work with texts on various topics, but when doing research on a specific topic, word lemmatizers become a more flexible tool. Preliminarily, the lemmatizer dictionary was reduced using INNER JOIN with a list of unique terms in the 'KEYWORDS' field. Further lemmatization of keywords was performed using sed utility and the created lemma dictionary.

In a common case the better option is the lemmatizer proposed by Krovetz, which implements dictionary lemmatization first and uses rule-based lemmatization for non-dictionary terms [7]. But in our case, there was a fixed list of keywords, and this approach felt redundant.

Keyword co-occurrence estimation was performed using the FP-growth utility developed by Christian Borgelt[2]. The scores obtained using this tool were compared with the results of a direct search for matched co-occurrence of terms in the 'KEYWORDS' field using the grep and wc utilities.

Keyword clustering and visualizing was done using the following programs:

- Scimago Graphica [8] which implements the Clauset-Newman-Moore algorithm [9], applicable to both weighted and unweighted graphs.
- The Agglomerative Hierarchical Clustering method, implemented in Multidendrograms, was employed to construct the dendrogram of Keywords [10].

Note: In this article, "term" is often used as a synonym for "keyword" and "score" is used as a synonym for specific parameters such as "minimum support".

Results and discussions

Dendrogram Construction Using FP Growth Algorithm Estimates and Direct Search Matches

To apply the agglomerative hierarchical clustering algorithm, keywords records were used to which lemmatization was applied and abbreviations given in parentheses for the terms used were removed. This is due to the fact that in some records only keywords were given, while in others they had abbreviations in parentheses.

The evaluation of joint occurrence of keywords was carried out with the following parameters of the utility: fpgrowth -s0.1m2n2. That is, the minimum support was equal to 0.1 and only the co-occurrence of two keywords was determined. This resulted in 39 pairs of keywords, which are listed in Table 1.

Table 1. The list of keyword pairs with minimum support of 0.1.

The first term	The second term	Score	Occur
deep_learn	machine_learn	0.266421	26
energy_transition	renewable_energy	0.245927	24
solar_energy	renewable_energy	0.225433	22
artificial_intelligence	machine_learn	0.204939	20
sustainability	renewable_energy	0.204939	21
anaerobic_digestion	biogas	0.174198	17
machine_learn	renewable_energy	0.174198	17
sustainable_development	renewable_energy	0.174198	17
energy_management	microgrid	0.163951	16

[2] https://borgelt.net/fpgrowth.html

 doi:10.20944/preprints202406.1380.v1

photovoltaic	renewable_energy	0.163951	16
artificial_neural_network	machine_learn	0.153704	15
combustion	emission	0.153704	15
microgrid	renewable_energy	0.153704	15
smart_grid	microgrid	0.153704	15
energy_storage	renewable_energy	0.143457	14
wind_energy	solar_energy	0.143457	14
gasification	biomass	0.13321	14
phase_change_material	thermal_energy_storage	0.13321	13
pyrolysis	biomass	0.13321	15
combustion	biomass	0.122963	13
energy_efficiency	renewable_energy	0.122963	12
forecast	machine_learn	0.122963	13
neural_network	machine_learn	0.122963	12
optimization	renewable_energy	0.122963	12
random_forest	machine_learn	0.122963	12
vehicle-to-grid	electric_vehicle	0.122963	12
combustion	hydrogen	0.112716	12
energy_policy	renewable_energy	0.112716	11
smart_grid	machine_learn	0.112716	11
solar_energy	photovoltaic	0.112716	12
sustainability	energy_efficiency	0.112716	11
wind_energy	renewable_energy	0.112716	11
biochar	pyrolysis	0.10247	10
energy_consumption	energy_efficiency	0.10247	10
energy_consumption	machine_learn	0.10247	10
energy_management_system	microgrid	0.10247	11
hydrogen	renewable_energy	0.10247	11
smart_grid	renewable_energy	0.10247	10
wind_energy	wind_turbine	0.10247	10

Where “Score” is the score obtained by the FP-growth algorithm; “Occur” (occurrence) is the number of matches of term pairs obtained by direct search.

Different values of co-occurrence of terms obtained by direct search with the same values given by the FP-growth algorithm are highlighted in yellow.

The first three columns of Table 1 were used to construct the dendrogram shown in Figure 1. MultiDendrograms-5.2.1 program parameters in use: Type of measure — Similarity; Precision — 6; Clustering algorithm — Arithmetic Linkage.

 doi:10.20944/preprints202406.1380.v1

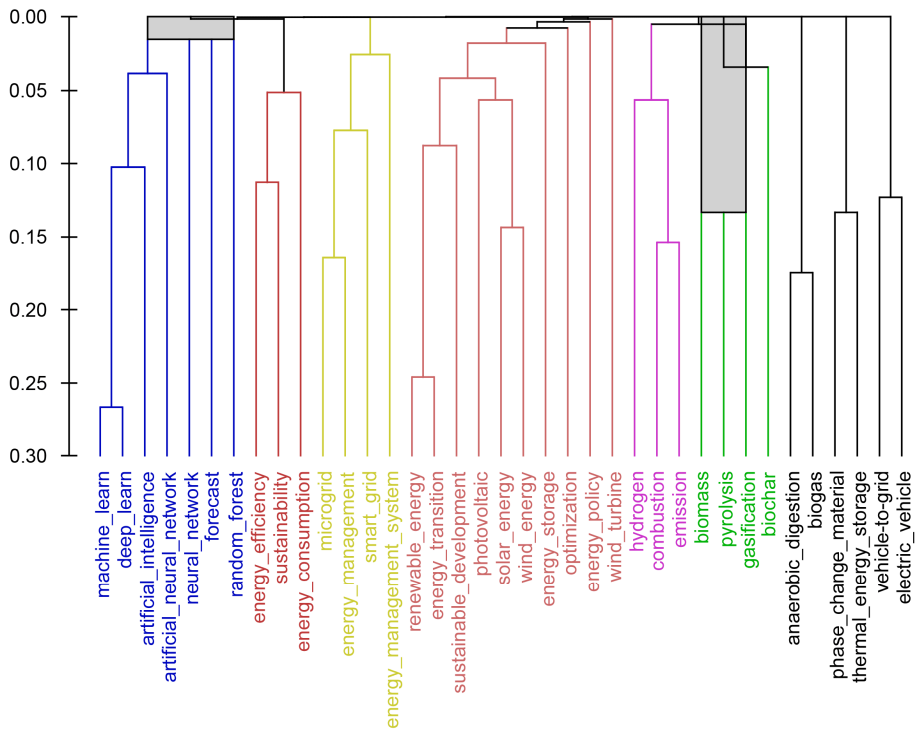

Figure 1. Dendrogram of co-occurrence of keywords obtained using the estimation given by the FP-growth algorithm.

The dendrogram constructed using the values obtained by direct searching is shown in Figure 2.

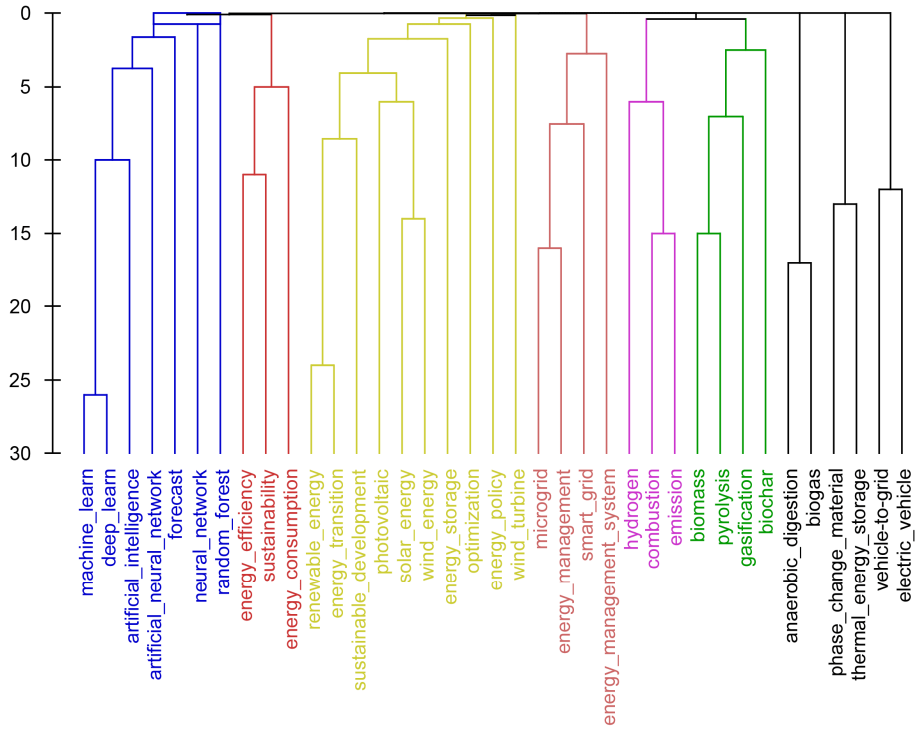

Figure 2. Dendrogram of keyword co-occurrence obtained using values from direct searches.

With the same parameters of the MultiDendrograms program, the dendrograms obtained using the FP-growth algorithm and direct search for keyword co-occurrence are practically the same.

Thus, in the context of this clustering approach, using the FP-growth algorithm gives good results.

 doi:10.20944/preprints202406.1380.v1

The advantage of dendrograms is that they are easy to interpret; for example, one can assume that the machine learning topic is closer to the energy efficiency topic than to other clusters, and that the hydrogen and biomass topics may have close objectives.

Using Keyword Co-Occurrence for Graph-Based Clustering

In the field of bibliometrics, programs for keyword clustering using graph algorithms are widely used. The most commonly used programs are VOSviewer and Bibliometrix. Request: 'sciencedirect.com/search?qs=VOSviewer' gives 4,766 results, and the request 'sciencedirect.com/search?qs=Bibliometrix' — 1,131. The first program uses the Leiden clustering algorithm [11], and in the second one — Louvain [12]. Both programs are well suited for routine bibliometric studies, but our study aims to compare two estimates of co-occurrence of keywords - given by the FP-growth algorithm and direct search for a small set of co-occurring terms.

Graphs 3 and 4 were constructed based on the data presented in Table 1, i.e., the same data used to construct the dendrograms presented in Figures 1 and 2.

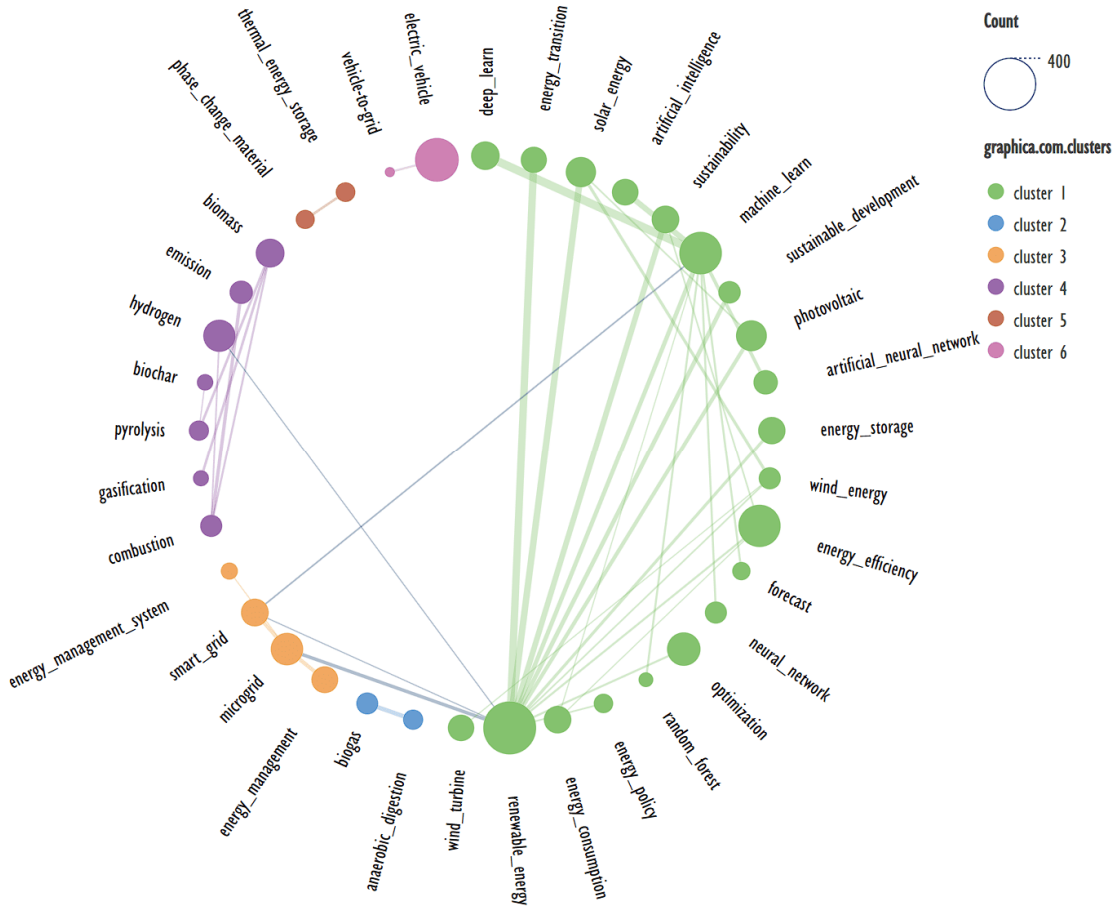

Figure 3. Clustering of keywords based on their co-occurrence when using the FP-growth utility score.

Contrary to dendrograms, in this case, the graphs reflecting keyword clustering differ more clearly when the parameters are equal. Figure 4 shows that "renewable energy" and "machine leaning" are placed in different clusters. Conversely, "microgrid" is placed in the same cluster as "renewable energy", while in Figure 3 it is in a separate cluster. It is also interesting that "energy consumption" is placed in the same cluster as "machine leaning".

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2024 doi:10.20944/preprints202406.1380.v1

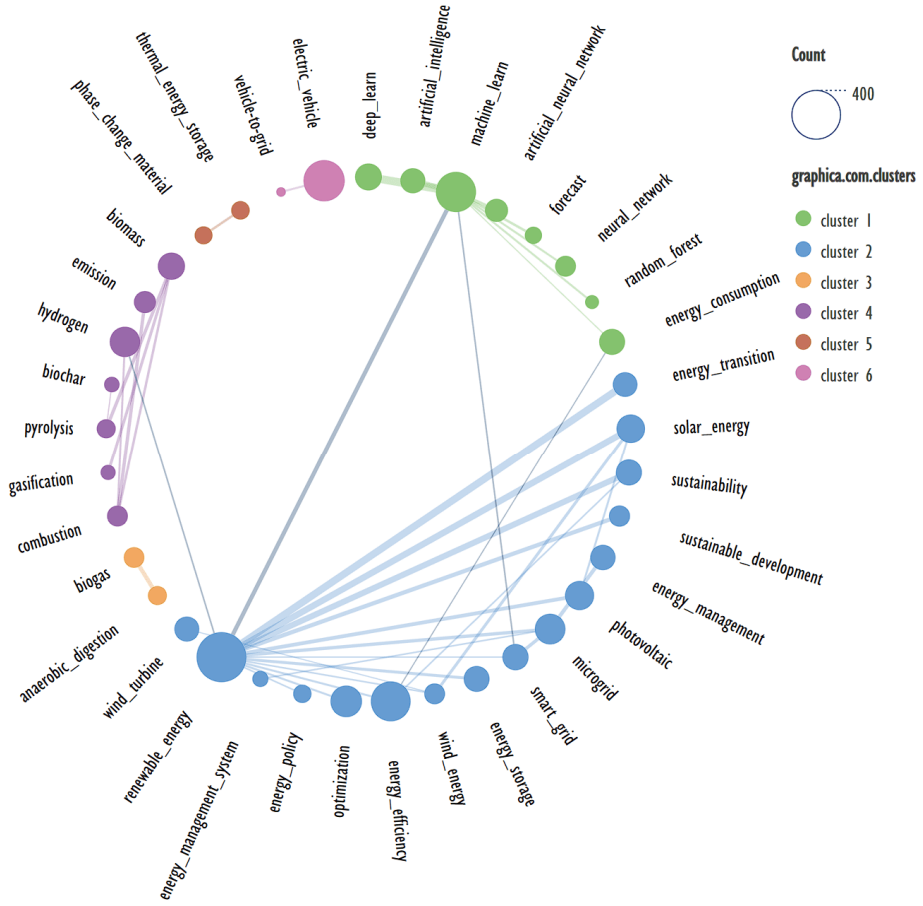

Figure 4. Clustering of keywords based on their co-occurrence when using direct search count.

Term clustering is quite sensitive to changes in the data being analyzed. Thus, it seems reasonable not to evaluate which variant is better, the estimation of co-occurrence represented by the FP-growth algorithm or direct search, but to analyze what causes the difference in the obtained results. In the given case, the circle diagrams are very clear, so in Figure 4 it can be seen that "renewable energy" and "machine leaning" have a close linkage. At the same time, "smart grid" has a rather weak link with "renewable energy", and with "microgrid" has a large one, so a small change in the analyzed data "smart grid" can be separated into a separate cluster with "microgrid".

Keyword clustering and visualization provide a framework for thinking about how to better shape queries to abstract databases, such as for systematic reviews [13], rather than focusing only on the formal side of clustering.

Construct an Alluvial Diagram Using the FP-Growth Algorithm

The co-occurrence of keywords can be used to construct an Alluvial diagram, see Figure 5.

In this case, we used data obtained when the minimum support parameter was set to 0.04 for the FP-growth algorithm. From the resulting set of keyword pairs, keywords/terms relevant to machine learning and numerical modeling (Term1) and their conjugates (Term2) are selected for graphing. Based on their interests, the subject matter expert can create his own sample of keywords, for example, among Term2 are widely represented terms: smart_grid, microgrid, energy_management, on them it is possible to build a diagram similar to the one shown in Figure 5.

 doi:10.20944/preprints202406.1380.v1

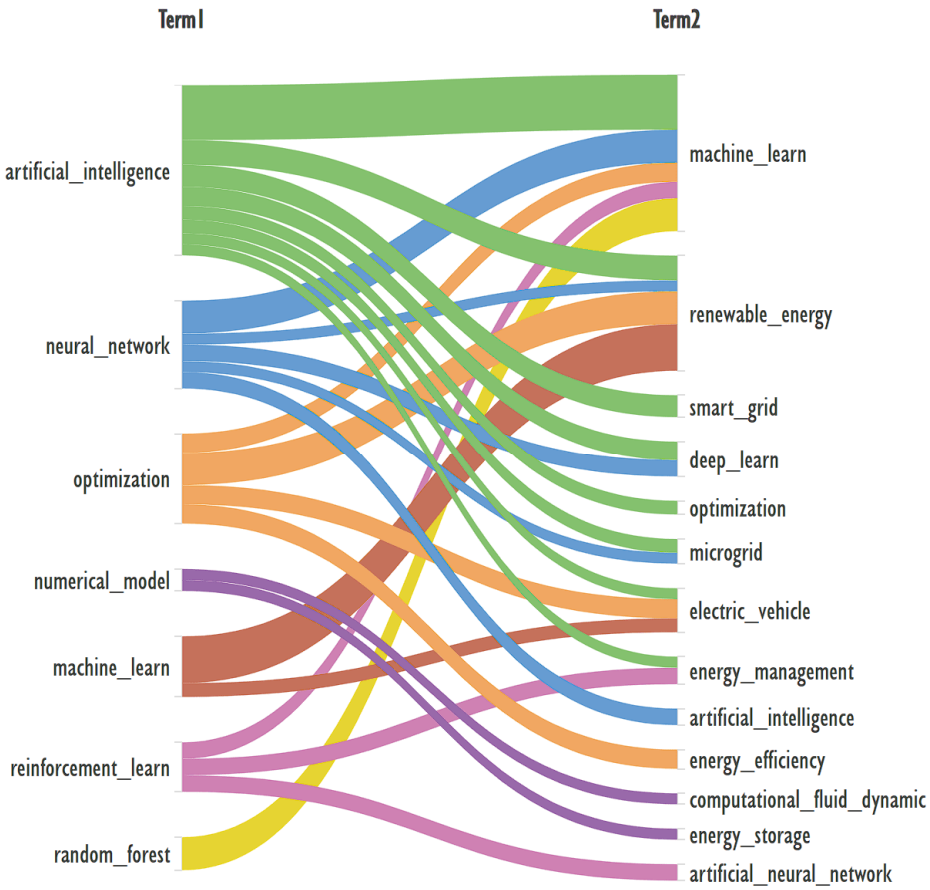

Figure 5. Alluvial diagram constructed for the co-occurrence of keywords that can be attributed to the topic of machine learning and numerical modeling.

The diagram in the figure clearly allows you to select the keyword pairs necessary for the query. For example, the topics reinforcement_learn energy_management or neural_network microgrid might be of interest.

At the used minimum support equal to 0.04, the number of records of keyword pairs was 462. Using the co-occurrence of three keywords, with the same support value, the number of variants decreases significantly and becomes equal to 28 and only 15 cases of co-occurrence of the four keywords.

For such a small number of records, the Alluvial diagram can be constructed in its entirety and is shown in Figure 6.

 doi:10.20944/preprints202406.1380.v1

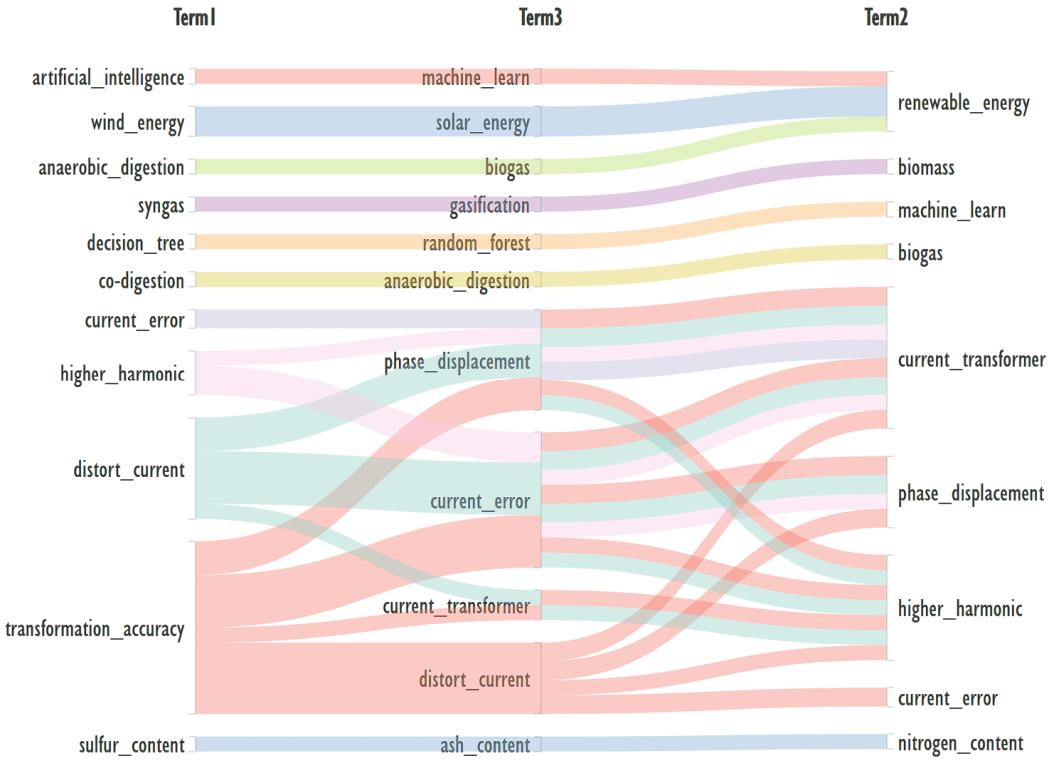

Figure 6. Alluvial diagram plotted for the co-occurrence of three keywords.

The combination of three keywords significantly narrows down the topics of the publications found by them and allows you to formulate an appropriate query.

Figure 6 shows that an interesting, niche topic can be defined by keywords and their combinations: higher_harmonic, phase_displacement, current_error, current_transformer, distort_current, transformation_accuracy.

Here are a few examples of publications that reflect this theme: [14–17].

Figures 5 and 6 are plotted using data with corresponding minimum support equal to 0.04. This is sufficient for plotting keyword pairs, but severely narrows the sample from the co-occurrence of three keywords.

To expand the data of records that contain keywords — current_error, current_transformer, distort_current, distort_voltage, higher_harmonic, instrument_transformer, phase_displacement, self-generation, transformation_accuracy, voltage_error, voltage_transformer co-occurrence of three keywords was determined when the minimum support of the FP-growth algorithm was reduced to 0.02. Figure 7 shows the result.

 doi:10.20944/preprints202406.1380.v1

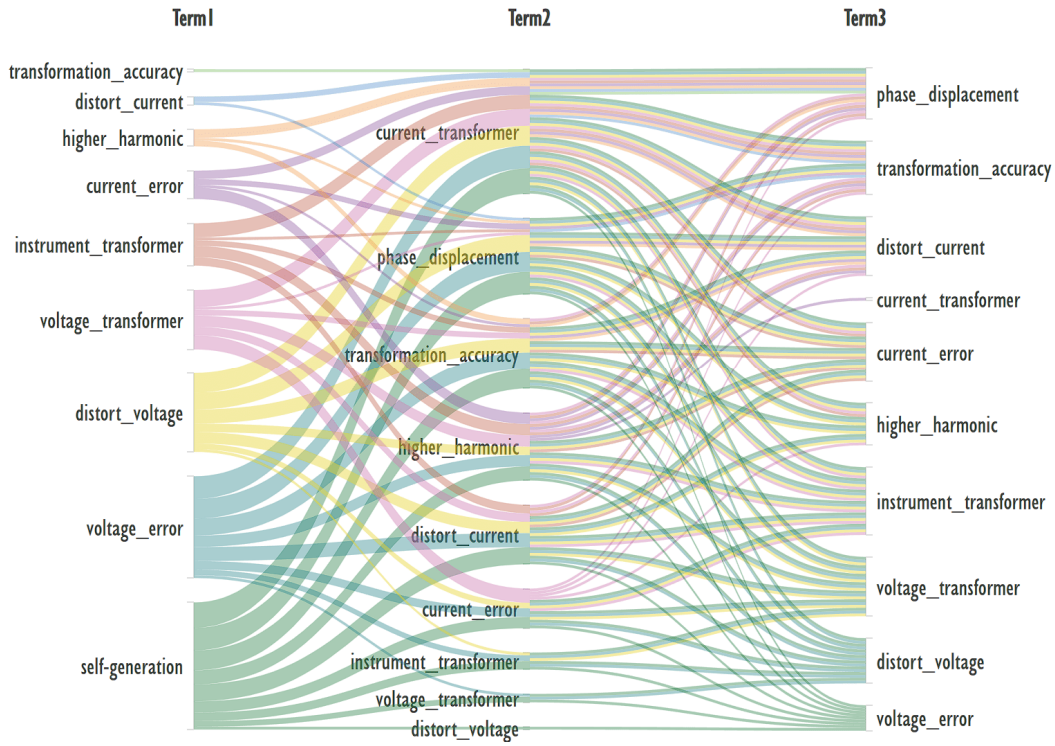

Figure 7. Alluvial diagram constructed for the co-occurrence of three keywords that satisfy the aforementioned condition.

In the context of topic-based querying, the value of this chart may be that for the most relevant topics in column Term1 - self-generation, voltage_error and distort_voltage - additional keywords presented in columns Term2 and Term3 can be easily matched to the most relevant ones.

Conclusions

The FP-growth algorithm provides estimates of keyword co-occurrence that are closely aligned with their direct search results. It can be employed at the stage of data preparation for subsequent clustering.

It is of interest to compare the keyword co-occurrence estimates obtained using the FP-growth algorithm and direct search in order to identify related but self-contained topics.

The co-occurrence of three or more keywords significantly reduces the number of potential keyword combinations for a given minimum support, thereby facilitating the identification of research topics that may be of interest but are narrow in scope.

A review of Energies bibliometric data for 2023-2024 shows that such a narrow prospective research topic can be described in the following terms: current_error, current_transformer, distort_current, distort_voltage, higher_harmonic, instrument_transformer, phase_displacement, self-generation, transformation_accuracy, voltage_error, voltage_transformer.

Possible Applications of the Findings

The results of this study can be used as a framework for compiling queries to abstract databases for the purpose of design and writing literature and systematic reviews.

Acknowledgment: This work was funded by the Ministry of Science and Higher Education of the Russian Federation, State Assignment No. 122022800270-0.

References

1. J. Han, J. Pei, and Y. Yin, 'Mining frequent patterns without candidate generation', *SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, Jun. 2000, doi: 10.1145/335191.335372.
2. C. Borgelt, 'An implementation of the FP-growth algorithm', in *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, Chicago Illinois: ACM, Aug. 2005, pp. 1–5. doi: 10.1145/1133905.1133907.
3. A. Augello, I. Infantino, G. Pilato, and F. Vella, 'Sensing the Web for Induction of Association Rules and their Composition through Ensemble Techniques', *Procedia Computer Science*, vol. 169, pp. 851–859, 2020, doi: 10.1016/j.procs.2020.02.152.
4. S. P. Singh, A. Sharma, and R. Kumar, 'Designing of fog based FBCMI2E Model using machine learning approaches for intelligent communication systems', *Computer Communications*, vol. 163, pp. 65–83, Nov. 2020, doi: 10.1016/j.comcom.2020.09.005.
5. K. Gadia and K. Bhowmick, 'Parallel Text Mining in Multicore Systems Using FP-tree Algorithm', *Procedia Computer Science*, vol. 45, pp. 111–117, 2015, doi: 10.1016/j.procs.2015.03.100.
6. H. Sun, B. Li, and B. Han, 'A novel keyphrase extraction method by combining FP-growth and LDA', in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Guilin: IEEE, Jul. 2017, pp. 1764–1768. doi: 10.1109/FSKD.2017.8393033.
7. R. Krovetz, 'Viewing morphology as an inference process', in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '93*, Pittsburgh, Pennsylvania, United States: ACM Press, 1993, pp. 191–202. doi: 10.1145/160688.160718.
8. Y. Hassan-Montero, F. De-Moya-Anegón, and V. P. Guerrero-Bote, 'SCImago Graphica: a new tool for exploring and visually communicating data', *EPI*, p. e310502, Sep. 2022, doi: 10.3145/epi.2022.sep.02.
9. A. Clauset, M. E. J. Newman, and C. Moore, 'Finding community structure in very large networks', *Phys. Rev. E*, vol. 70, no. 6, p. 066111, Dec. 2004, doi: 10.1103/PhysRevE.70.066111.
10. A. Fernández and S. Gómez, 'Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms', *J Classif*, vol. 25, no. 1, pp. 43–65, Jun. 2008, doi: 10.1007/s00357-008-9004-x.
11. V. A. Traag, L. Waltman, and N. J. Van Eck, 'From Louvain to Leiden: guaranteeing well-connected communities', *Sci Rep*, vol. 9, no. 1, p. 5233, Mar. 2019, doi: 10.1038/s41598-019-41695-z.
12. M. Aria and C. Cuccurullo, 'bibliometrix : An R-tool for comprehensive science mapping analysis', *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.
13. K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, 'Five Steps to Conducting a Systematic Review', *J R Soc Med*, vol. 96, no. 3, pp. 118–121, Mar. 2003, doi: 10.1177/014107680309600304.
14. M. Kaczmarek and E. Stano, 'New Approach to Evaluate the Transformation Accuracy of Inductive CTs for Distorted Current', *Energies*, vol. 16, no. 7, p. 3026, Mar. 2023, doi: 10.3390/en16073026.
15. M. Kaczmarek and E. Stano, 'Challenges of Accurate Measurement of Distorted Current and Voltage in the Power Grid by Conventional Instrument Transformers', *Energies*, vol. 16, no. 6, p. 2648, Mar. 2023, doi: 10.3390/en16062648.
16. M. S. Ballal, M. G. Wath, and H. M. Suryawanshi, 'A Novel Approach for the Error Correction of CT in the Presence of Harmonic Distortion', *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 4015–4027, Oct. 2019, doi: 10.1109/TIM.2018.2884575.
17. E. Stano and M. Kaczmarek, 'Analytical method to determine the values of current error and phase displacement of inductive current transformers during transformation of distorted currents higher harmonics', *Measurement*, vol. 200, p. 111664, Aug. 2022, doi: 10.1016/j.measurement.2022.111664.