

Article

Not peer-reviewed version

GSDMM Clustering Results Visualization Technique for Short Texts

[Boris Chigarev](#)*

Posted Date: 9 December 2024

doi: 10.20944/preprints202412.0705.v1

Keywords: short texts; GSDMM algorithm; visualization technique; bibliometric records; Scopus; VOSviewer; Scimago Graphica



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

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

Article

GSDMM Clustering Results Visualization Technique for Short Texts

Boris Chigarev

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

Abstract: The aim of the study is to propose a technique for visualizing the results of short text clustering using the GSDMM algorithm, in order to facilitate the analysis of the results and the selection of the hyperparameters of this algorithm and dictionary. GSDMM is selected as the most popular short text clustering algorithm on GITHUB. The algorithm implemented by Ryan Walker on Rust was used. The program Scimago Graphica was used to create bar charts. 16486 bibliometric records on the topic "Visualization" exported from the Scopus database on November 12, 2024 served as the source of short texts. Only Author keywords are used as short texts in this paper. Results. A technique for visualizing the results of short text clustering using the GSDMM algorithm is proposed, which is based on comparing the occurrence of keywords in a given cluster and in each of the other clusters. It is shown that the cluster topics obtained using the GSDMM algorithm can be compared with the results of author keyword clustering performed using the VOSviewer program. The obtained results can be interpreted as a certain stability of cluster themes obtained by essentially different methods. The author suggests to expand the study by creating a thematic dictionary of abbreviations, analyzing the influence of the dictionary on the clustering results of the GSDMM algorithm, and extending the method of visualizing the clustering results to other short texts such as titles and abstracts.

Keywords: short texts; GSDMM algorithm; visualization technique; bibliometric records; Scopus; VOSviewer; Scimago Graphica

Introduction

The **objective of the study** is to propose a technique for visualizing the results of short text clustering using the GSDMM algorithm to facilitate the analysis of the results and the selection of hyperparameters of this algorithm and dictionary.

Why GSDMM? It is the most popular short text clustering algorithm on GITHUB. Additionally, a query on ScienceDirect for "Title, Abstract, Keywords: 'short text' clustering algorithm" returns 27 results across all years. In the exported RIS file of these 27 records, GSDMM appears 12 times in 6 papers, LDA appears 17 times in 6 papers, "suffix tree" appears 4 times in one paper. But GSDMM is close to LDA, but for short texts — one topic per text.

BERTopic is also widely used for studying text topics, but it is not an algorithm, but a whole sequence of procedures. According to the developer of the program "By default, the main steps for topic modeling with BERTopic are sentence-transformers, UMAP, HDBSCAN, and c-TF-IDF run in sequence."¹ Even more, UMAP² — procedure using its own parameters [1], and "sentence-transformers" is also dependent on the training sample. BERT is suitable for vectorizing short texts on Twitter, but other models may be more appropriate for different text sources. [2]. Even if BERTopic gives good results, the analysis of the influence of the parameters of the procedures at each step on the final results seems questionable, so this program was excluded as an alternative to GSDMM.

¹ <https://github.com/MaartenGr/BERTopic> — BERTopic is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters.

² <https://github.com/lmcinnes/umap> — Uniform Manifold Approximation and Projection

Choosing a **GSDMM implementation**. According to the query “<https://github.com/search?q=short%20text%20clustering&type=repositories>”, the algorithm GSDMM: Short Text Clustering implemented by Ryan Walker has the highest rating and is implemented in both Python and Rust. It should be noted that there are also implementations of this algorithm in Java and GO.

Visualization of clustering results using the GSDMM algorithm. A query on GITHUB “GSDMM” returns 46 outputs, but the visualization of the results of this algorithm is only implemented in two Python projects: “center-for-humanities-computing/tweetopic”³ and “Illias-b/Natural-Language-Processing”⁴. In both cases, the results are presented in the form of term clouds and bar charts.

Note: The Center for Humanities Computing is notable for publishing the journal Computational Humanities Research, showcasing cutting-edge research at the intersection of computational methods and the humanities.

The advantage of the two visualization approaches described above is that they can be done quickly. A number of parameters can be set, but the implementation procedures remain the same. This advantage is also the main disadvantage — lack of flexibility in the choice of procedures at each step, especially in text preparation, dictionary selection and bar chart construction.

What is good for constant monotonous work is not always suitable for research work, where, in my opinion, it is more appropriate to use the well-known approach: the program does one thing, but does it well. In this case, at each stage it is easier to analyze the result, adjust parameters and choose procedures. The work takes more time, but it allows you to analyze the whole process in more depth. For example, in this study, the bar chart is displayed not by the ratio of the occurrence of a term in one cluster to its total occurrence, but by the difference between the occurrence of a term in one particular cluster and other individual clusters. The flexible text preprocessing and dictionary selection for GSDMM were also used.

The novelty of this study lies in the choice of bar chart for visualizing the results of text clustering using GSDMM and comparing the topics of clusters with the clustering based on the co-occurrence of keywords.

The following sections discuss the detailed implementation of the above.

Materials and Methods

The source of the short texts was bibliometric records on the “Visualization” topic exported from Scopus, the scientific abstracts and citations database on November 12, 2024.

In order to focus on the issue of visual presentation of clustering results, this paper only uses the Author keywords as short texts. The use of titles and abstracts as short texts requires a more detailed choice of their preprocessing and is not considered in this paper.

Bibliometric data were exported on request: (TITLE-ABS-KEY(“Visual Analytics” OR “Data Visualization” OR “Information Visualization” OR “Visualization Technique” OR “Visual Representations” OR “Visual Languages” OR “Visualization Tools” OR “Visual Exploration” OR “Visualization System” OR “Visual Analysis” OR “Interactive Visualizations” OR “Flow Visualization”) AND (LIMIT-TO (DOCTYPE,“ar”)) AND (LIMIT-TO (LANGUAGE,“English”)) AND (LIMIT-TO (SUBJAREA,“ENGI”) OR LIMIT-TO (SUBJAREA,“COMP”) OR LIMIT-TO (SUBJAREA,“ENER”)) AND (LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR,2021) OR LIMIT-TO (PUBYEAR,2022) OR LIMIT-TO (PUBYEAR,2023) OR LIMIT-TO (PUBYEAR,2024))). 16486 records were received.

³ <https://github.com/centre-for-humanities-computing/tweetopic> — Blazing fast topic modelling for short texts

⁴ <https://github.com/Illias-b/Natural-Language-Processing/> — A concise project analysing NPS survey data using Python tools and GSDMM for topic modelling

“visualization” occurs in 8 out of 9 clusters, so the size of this word in the figure is proportional to its occurrence in each cluster. It is typical for document clustering that keywords describing the subject of a cluster can occur in multiple clusters.

These are only the terms included in the 146 records analyzed. A more complete inclusion of keywords is hindered by the readability of the figure in the paper. Choosing the best graphical representation of the term cloud was not the goal of this paper.

Analyzing the Data of Each Cluster Obtained Using the GSDMM Algorithm

The GSDMM utility generates a file ‘prefix-cluster_descriptions.txt’, which contains for each cluster a list of vocabulary words and their occurrences. The cluster numbers are kept in the original sequence: the first non-empty cluster is labeled as 5, the next cluster as 14, and so on. In addition, the GSDMM utility generates a file indicating the cluster to which each entry/document is assigned and a probability score indicating the degree of correspondence between the entry and the cluster. However, in this particular study, the aim was to select a suitable visual representation of the cluster descriptions based on dictionary words. Therefore, this file was not used in this paper.

To implement visualization, the keywords (words from the dictionary) in a separate cluster, and the other clusters were compared. The data were presented graphically in Figures 2-10.

The archive contains HTML pages with interactive charts that facilitate a fuller understanding of the results.

The occurrence of keywords was compared using the left join operation of this cluster and others. Then the frequency of keywords in this cluster and other clusters was subtracted.

Terms that occurred in the current cluster but did not occur in other clusters were marked as ‘0’ cluster in the graphs to emphasize that they were not clusters other than the current cluster. If there were no such terms, this was labeled as ‘null’ in the graphs. For example, see Figures 2 and 3.

In order to limit the size of the graph, only keywords whose occurrence exceeded a certain threshold, which was different for each cluster, were displayed on the chart. This value was indicated in the caption below the figure.

To emphasize the importance of keywords that occur in the cluster in question but not in other clusters, they can be plotted even if their frequency is lower than the one specified and indicated in the figure caption. For example, Figure 4 shows terms that occur 39 or more times in cluster 15 and also occur in other clusters. However, terms that do not occur in other clusters are shown even if they occur only 10 or more times in cluster 15.

Figure 2 shows the difference between the occurrence of the most frequent keywords in Cluster 5 and the other clusters.

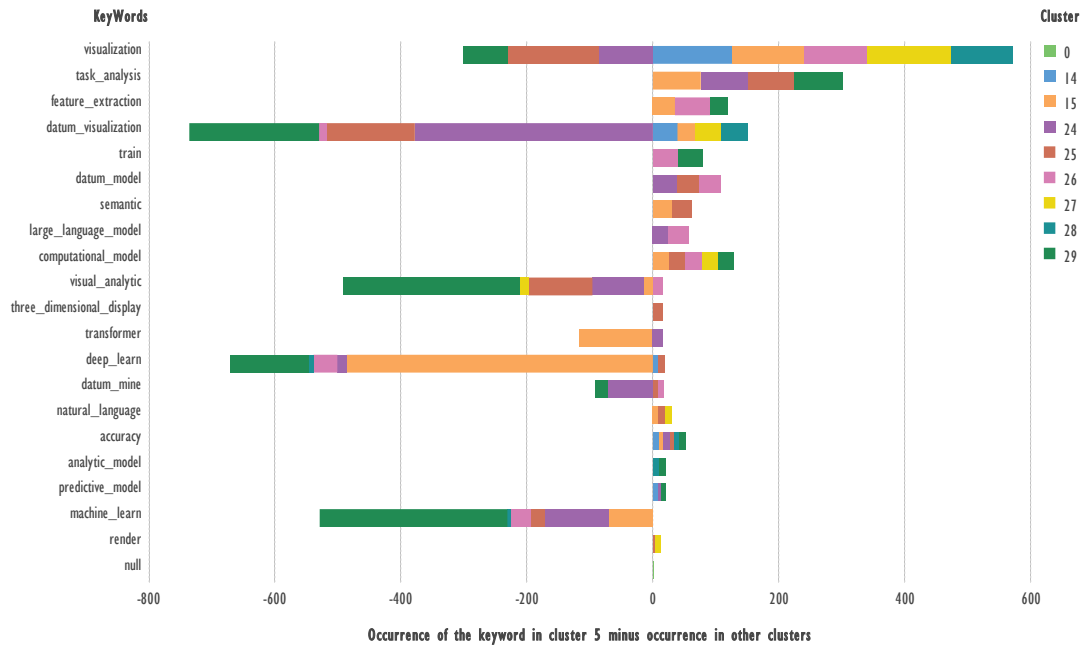


Figure 2. Comparison of keywords from Cluster 5 and the rest of the clusters. Terms with 10 or more occurrences were used.

There are no keywords in this cluster that do not occur in other clusters.

Terms: 'task_analysis', 'feature_extraction', 'train', 'datum_model', 'semantic', 'large_language_model', 'computational_model', which occur more frequently in cluster 5 than in other clusters, can serve as candidates to describe the cluster theme. 'visualization' occurs in many clusters. 'datum_visualization', 'visual_analytic', 'deep_learn', 'machine_learn' are more likely to occur in other clusters.

Figure 3 shows the difference between the occurrence of the most frequent keywords in Cluster 14 and the other clusters.

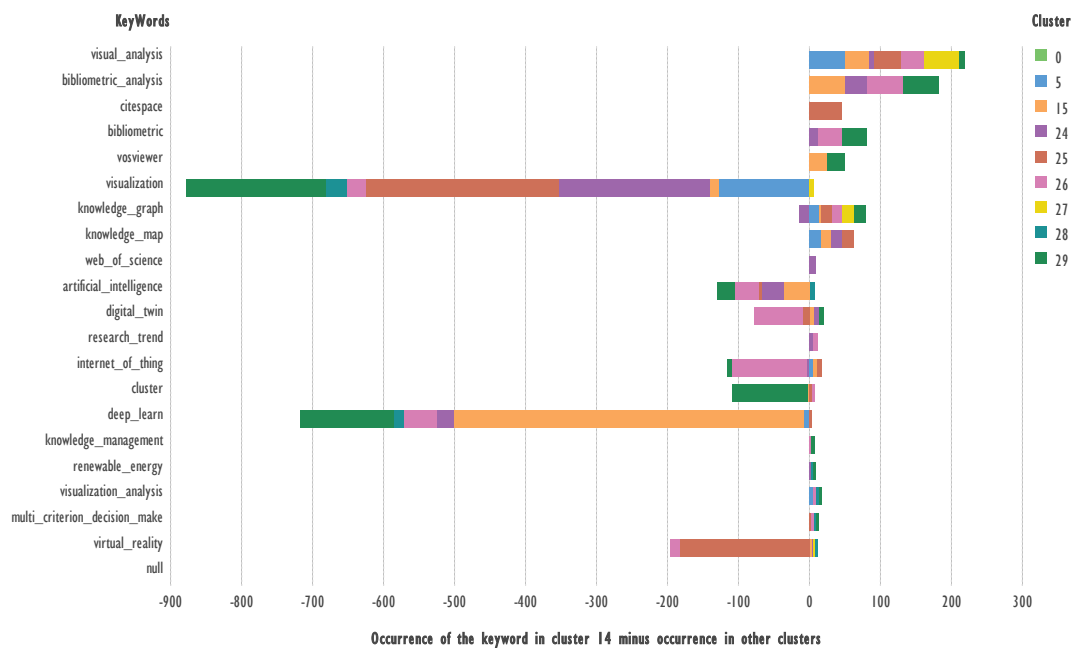


Figure 3. Comparison of keywords from Cluster 14 and the rest of the clusters. Terms with 5 or more occurrences were used.

There are no keywords in this cluster that do not occur in other clusters.

Terms: 'visual_analysis', 'bibliometric_analysis', 'citespace', 'bibliometric', 'vosviewer', 'knowledge_map', 'web_of_science' which occur more frequently in cluster 14 than in other clusters, can serve as candidates to describe the cluster theme; 'visualization', 'deep_learn', 'virtual_reality' are more likely to occur in other clusters.

Figure 4 shows the difference between the occurrence of the most frequent keywords in Cluster 15 and the other clusters.

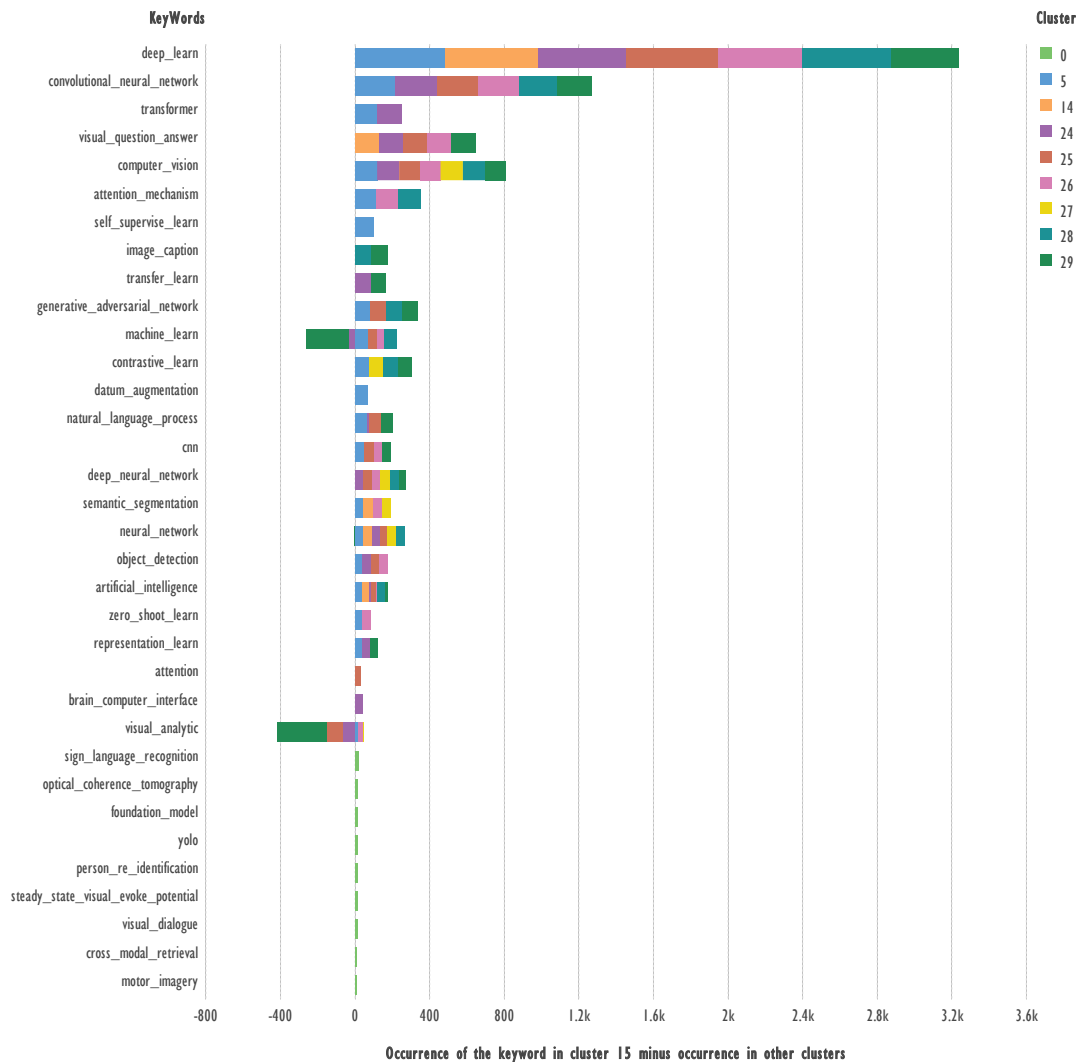


Figure 4. Comparison of keywords from Cluster 15 and the rest of the clusters. Terms with 39 or more occurrences were used.

Terms: 'sign_language_recognition → 19', 'optical_coherence_tomography → 16', 'foundation_model → 14', 'yolo → 14', 'person_re_identification → 13', 'steady_state_visual_evoke_potential → 13', 'visual_dialogue → 11', 'cross_modal_retrieval → 10', 'motor_imagery → 10' occurs less than 39 times, but added to the figure to make it more fully understood.

Terms: 'deep_learn', 'convolutional_neural_network', 'transformer', 'visual_question_answer', 'computer_vision', 'attention_mechanism', 'generative_adversarial_network', 'contrastive_learn' which occur more frequently in cluster 15 than in other clusters, can serve as candidates to describe the cluster theme; 'machine_learn', 'visual_analytic' are more likely to occur in other clusters.

Figure 5 shows the difference between the occurrence of the most frequent keywords in Cluster 24 and the other clusters.

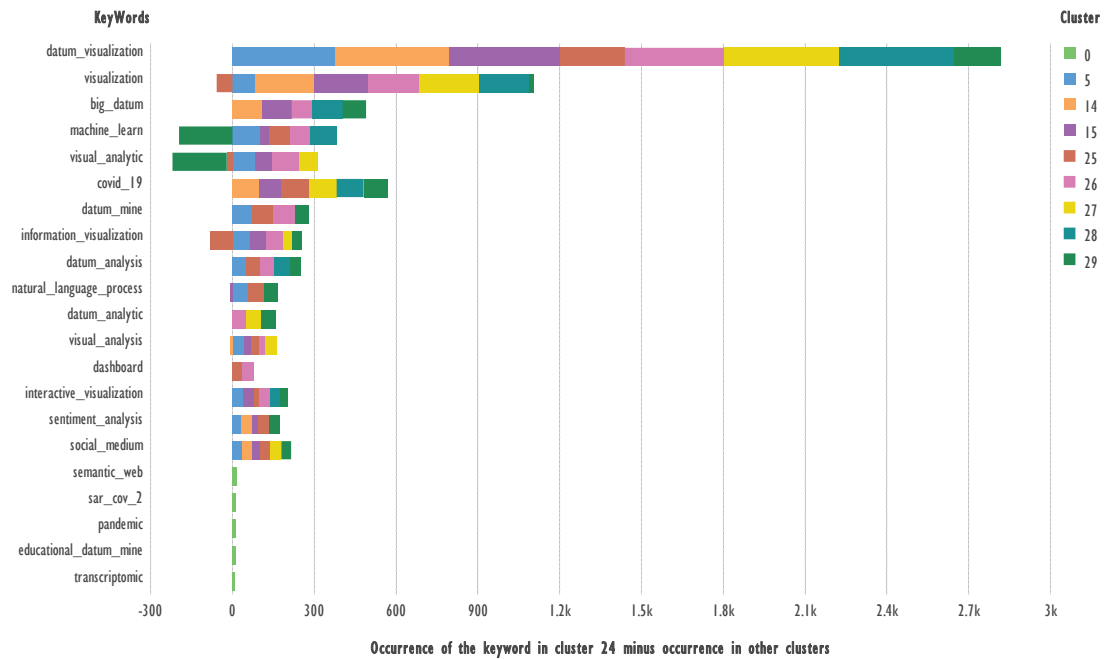


Figure 5. Comparison of keywords from Cluster 24 and the rest of the clusters. Terms with 40 or more occurrences were used.

Terms: 'semantic_web 15', 'sar_cov_2 14', 'pandemic 13', 'educational_datum_mine 12', 'transcriptomic 11' occurs less than 40 times, but added to the figure to make it more fully understood.

Terms: 'datum_visualization', 'big_datum', 'covid_19', 'datum_mine', 'datum_analysis', 'interactive_visualization', 'sentiment_analysis', 'sentiment_analysis', 'social_medium' which occur more frequently in cluster 24 than in other clusters, can serve as candidates to describe the cluster theme; 'machine_learn', 'visual_analytic' are more likely to occur in other clusters. Topic 'covid_19' has been analyzed extensively in scientific publications.

Figure 6 shows the difference between the occurrence of the most frequent keywords in Cluster 25 and the other clusters.

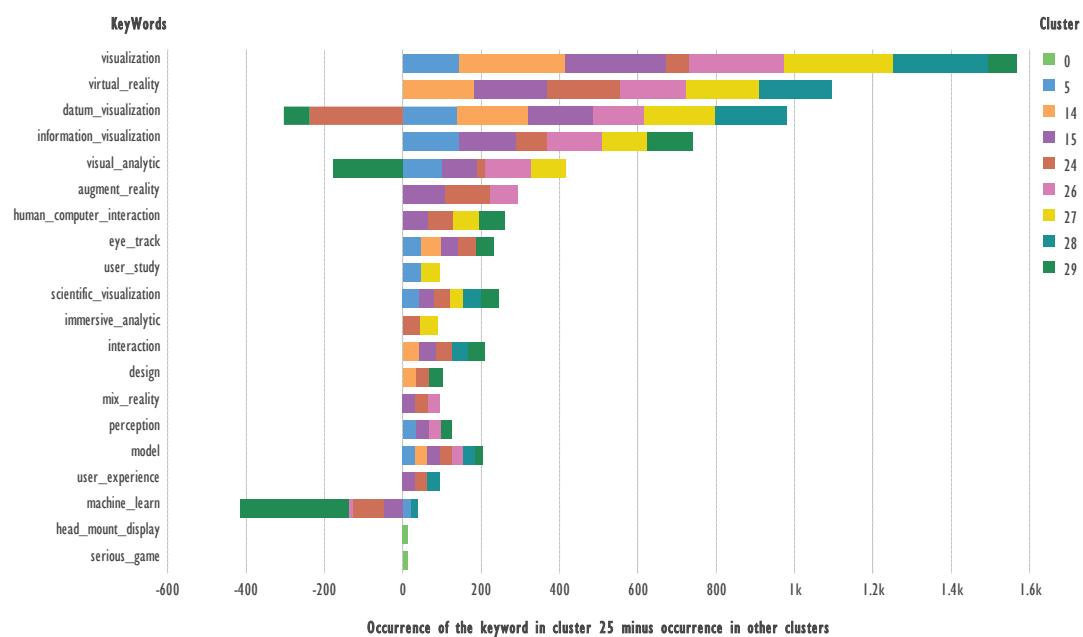


Figure 6. Comparison of keywords from Cluster 25 and the rest of the clusters. Terms with 32 or more occurrences were used.

Terms: 'head_mount_display 13', 'serious_game 12', occurs less than 32 times, but added to the figure to make it more fully understood.

Terms: 'visualization', 'virtual_reality', 'information_visualization', 'augment_reality', 'human_computer_interaction', 'eye_track', 'scientific_visualization', 'immersive_analytic', 'interaction' which occur more frequently in cluster 25 than in other clusters, can serve as candidates to describe the cluster theme; 'machine_learn' occur in other clusters.

Figure 7 shows the difference between the occurrence of the most frequent keywords in Cluster 26 and the other clusters.

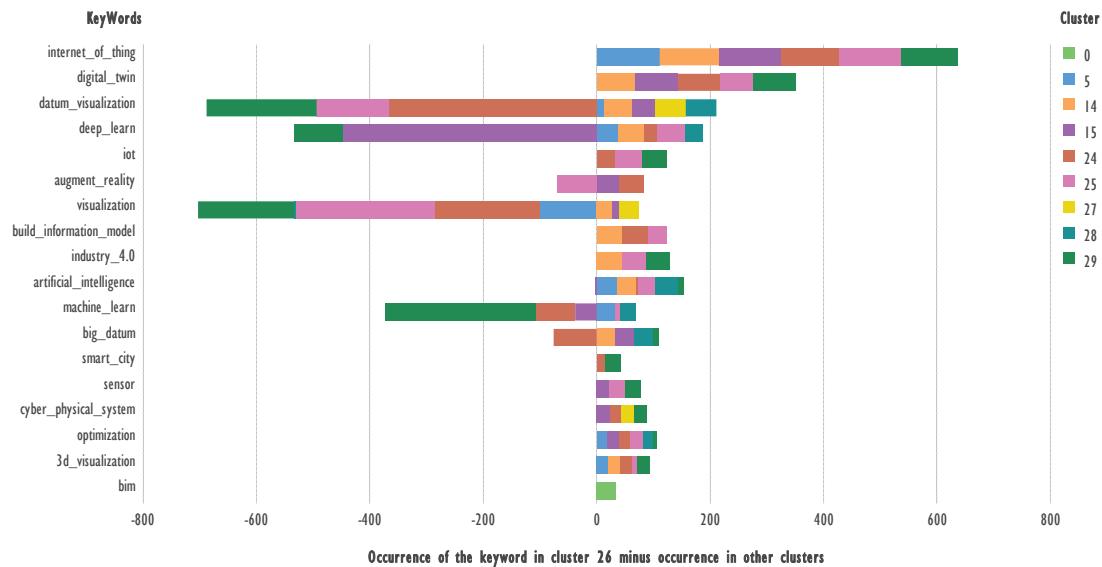


Figure 7. Comparison of keywords from Cluster 26 and the rest of the clusters. Terms with 23 or more occurrences were used.

Terms: 'internet_of_thing', 'digital_twin', 'iot', 'build_information_model', 'industry_4.0', 'artificial_intelligence', 'cyber_physical_system', 'optimization', '3d_visualization' which occur more frequently in cluster 26 than in other clusters, can serve as candidates to describe the cluster theme; 'datum_visualization', 'deep_learn', 'visualization', 'machine_learn' occur in other clusters.

Note: No abbreviation substitution was used in the preparation of the texts, so 'internet_of_thing' and 'iot' are marked as different terms. It is useful to compile a list of thematic abbreviations for use in future work.

Figure 8 shows the difference between the occurrence of the most frequent keywords in Cluster 27 and the other clusters.

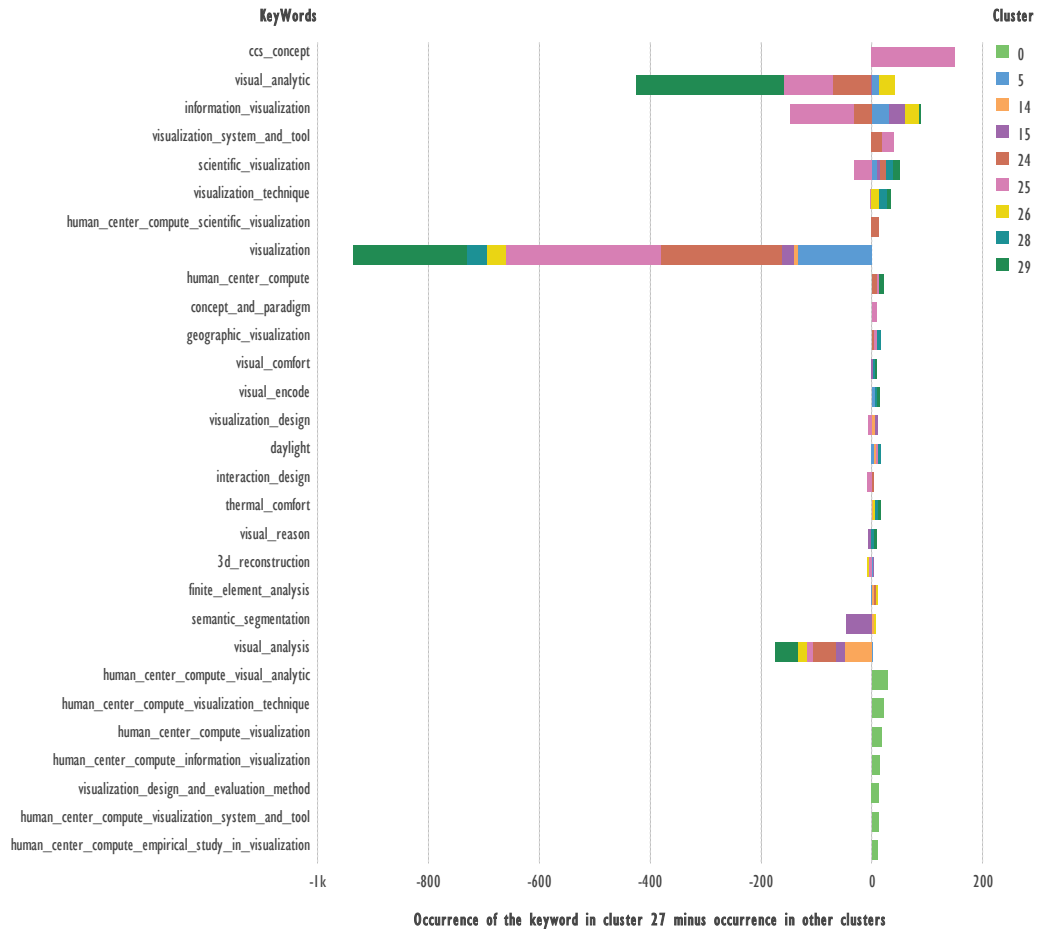


Figure 8. Comparison of keywords from Cluster 27 and the rest of the clusters. Terms with 4 or more occurrences were used.

Terms: 'human_center_compute_visual_analytic',
 'human_center_compute_visualization_technique', 'human_center_compute_visualization',
 'human_center_compute_information_visualization',
 'visualization_design_and_evaluation_method',
 'human_center_compute_visualization_system_and_tool',
 'human_center_compute_empirical_study_in_visualization' — occur only in cluster 27, reflecting the specificity of the cluster topic.

Terms: 'ccs_concept', 'information_visualization', 'visualization_system_and_tool',
 'visualization_technique', 'human_center_compute_scientific_visualization' which occur more frequently in cluster 14 than in other clusters, can serve as candidates to describe the cluster theme;
 'visual_analytic', 'visualization', 'visual_analysis' occur in other clusters.

Figure 9 shows the difference between the occurrence of the most frequent keywords in Cluster 28 and the other clusters.

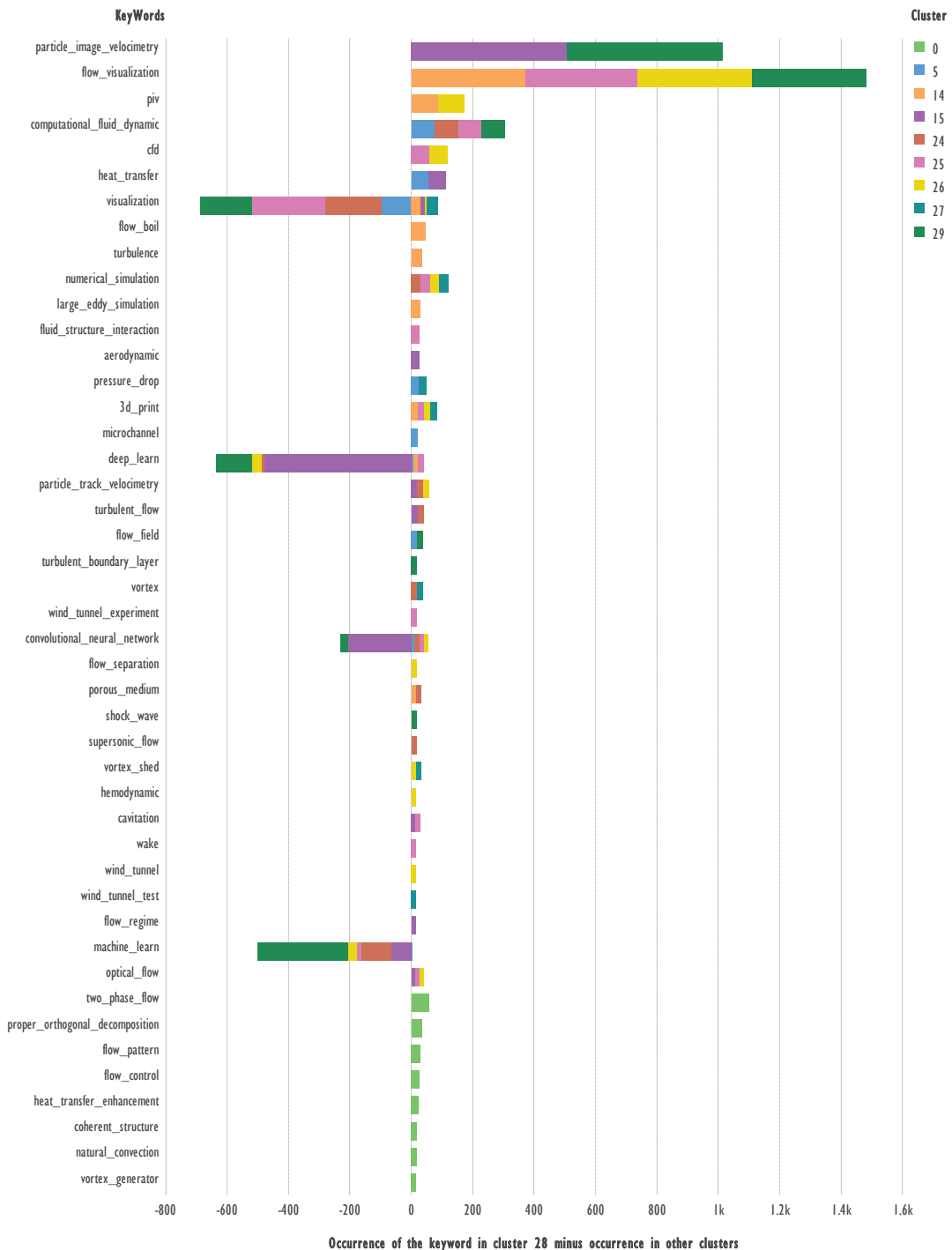


Figure 9. Comparison of keywords from Cluster 28 and the rest of the clusters. Terms with 15 or more occurrences were used.

Terms: 'two_phase_flow', 'proper_orthogonal_decomposition', 'flow_pattern', 'flow_control', 'heat_transfer_enhancement', 'coherent_structure', 'natural_convection', 'vortex_generator' — occur only in cluster 28, reflecting the specificity of the cluster topic.

Terms: 'particle_image_velocimetry', 'flow_visualization', 'piv', 'computational_fluid_dynamic', 'cid', 'heat_transfer', 'flow_boil', 'turbulence', 'numerical_simulation', '3d_print' which occur more frequently in cluster 28 than in other clusters, can serve as candidates to describe the cluster theme; 'visualization', 'deep_learn', 'convolutional_neural_network' occur in other clusters.

Figure 10 shows the difference between the occurrence of the most frequent keywords in Cluster 29 and the other clusters.

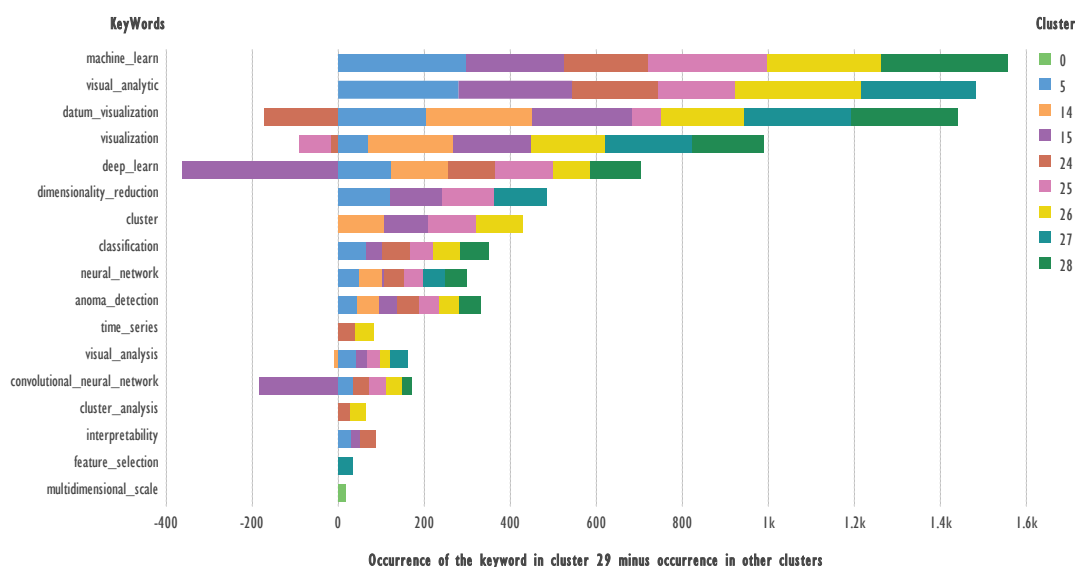


Figure 10. Comparison of keywords from Cluster 29 and the rest of the clusters. Terms with 34 or more occurrences were used.

Term: 'multidimensional_scale' — occurs only in cluster 29, reflecting the specificity of the cluster topic.

Terms: 'machine_learn', 'visual_analytic', 'dimensionality_reduction', 'cluster', 'classification', 'neural_network', 'anoma_detection', 'time_series', 'visual_analysis', 'cluster_analysis', 'interpretability', 'feature_selection' which occur more frequently in cluster 29 than in other clusters, can serve as candidates to describe the cluster theme; 'deep_learn', 'convolutional_neural_network' — frequently appear in cluster 15.

Comparison of Author Keywords in VOSviewer and GSDMM Clusters

In bibliometric analysis, keyword clustering using VOSviewer software is more commonly used rather than document clustering using GSDMM.

Thus, the simplest query to the open abstract database ScienceDirect 'document clustering gsdmm' gives only 40 publications, the query 'Gibbs Sampling' 'Dirichlet Multinomial Mixture' — 60 results, and the query 'keyword clustering VOSviewer' returns 4,570 results.

Therefore, it is interesting to compare the Author keywords clusters obtained with the VOSviewer program and the Author keywords specific to the record/document clusters obtained with the GSDMM algorithm. The approaches are quite different, but even an approximate overlap of dominant themes can indicate the semantic meaningfulness of the results obtained by the different methods.

For this comparison, the same prepared Author keywords records used in clustering with the GSDMM algorithm were imported into the VOSviewer software and a co-occurrence network of terms was constructed. The results are shown in Figure 11.

Using the default parameters, 13 clusters were obtained. In order to obtain an equal number of clusters to compare the results of the two approaches, the parameter reflecting the minimum number of keywords in a cluster was increased to 45.

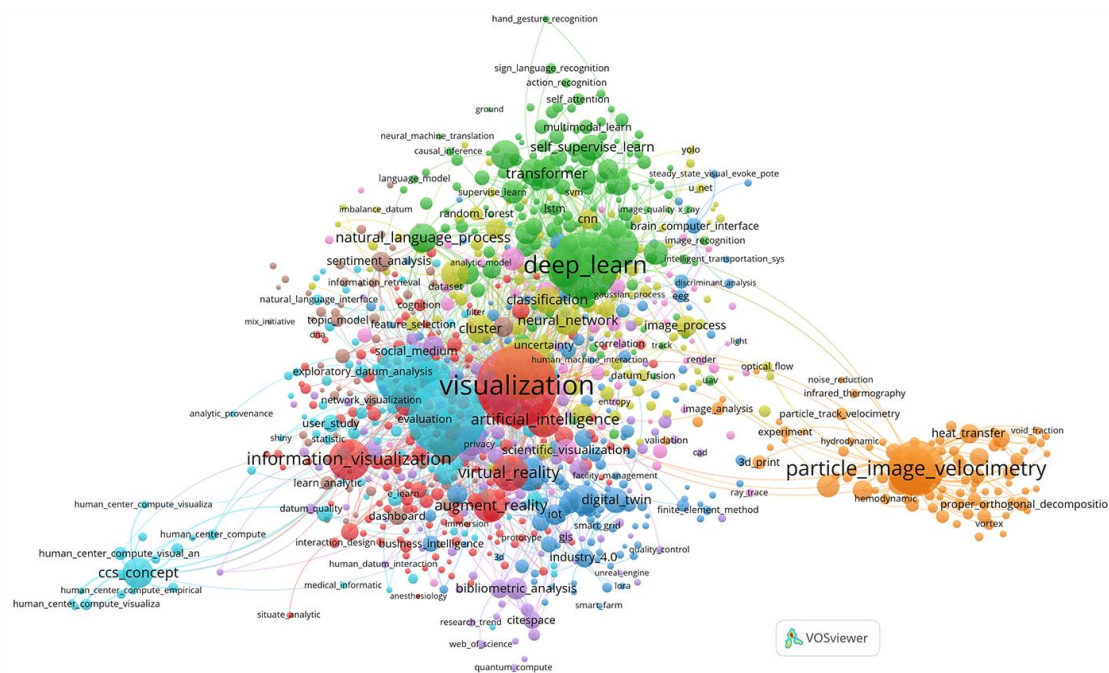


Figure 11. Network of 9 clusters of co-occurrences of Author keywords. Obtained with the minimum number of terms in the cluster equal to 45.

The orange cluster, which can be tentatively called “particle_image_velocimetry”, stands out well in this figure; the terms of this cluster are also typical of the 28 cluster shown in Figure 9.

For a detailed examination of the results shown in Figure 11, the file ‘AuKW16486-45-45-45.json’ included in the archive can be used at the service <https://app.vosviewer.com/>.

If we use the interactive features of the ‘app.vosviewer’ service, we can see that terms like ‘visualization’ or ‘deep_learn’ often co-occur with keywords from other clusters. Unlike document/record clustering, when clustering keywords, they belong to one cluster but may co-occur with keywords from other clusters.

The frequent occurrence of terms such as ‘visualization’ in different clusters of documents/records is quite understandable, as this term is included in the query for which data was collected in Scopus. Other common terms, such as ‘deep learning’, are not included in the query, but are often found in the broader context of papers about data.

The following is a brief comparison of keywords in clusters obtained using the GSDMM algorithm and the VOSviewer program. The aim is to find out the reproducibility of the identified topics obtained by significantly different methods.

The analysis was carried out as follows: for each cluster obtained by the two methods, a list of keywords and their occurrence was determined. From these lists, a sample of the 20 most frequent keywords was taken. The tables obtained were compared by INNER JOIN. The results of the comparison of the most similar tables are presented in Tables 1–10. Abbreviations in table titles: VOSviewer-AuKW — author keywords from VOSviewer cluster; GSDMM-AuKW — author keywords from GSDMM cluster. N — Number of Author Keywords in the cluster. VOSviewer Author Keywords can only belong to one cluster, while GSDMM Author Keywords can belong to several clusters, so they are not more than VOSviewer Author Keywords.

Table 1. INNER JOIN of top 20 terms from VOSviewer cluster 9 against top 20 terms from GSDMM cluster 5.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
task_analysis	89	task_analysis	78
datum_model	48	datum_model	39
train	44	train	41
computational_model	33	computational_model	27
three_dimensional_display	19	three_dimensional_display	17
natural_language	18	natural_language	12
render	18	render	10

This table summarizes typical “task_analysis” steps. An example of the work could be [5].

Table 2. INNER JOIN of top 20 terms from VOSviewer cluster 5 against top 20 terms from GSDMM cluster 14.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
visual_analysis	205	visual_analysis	53
bibliometric_analysis	78	bibliometric_analysis	52
knowledge_graph	71	knowledge_graph	17
bibliometric	60	bibliometric	35
citespace	52	citespace	49
vosviewer	28	vosviewer	26
knowledge_map	23	knowledge_map	17

This table shows: Visualization of bibliometric data is a relevant research topic. An example of the work could be [6].

Table 3. INNER JOIN of top 20 terms from VOSviewer cluster 2 against top 20 terms from GSDMM cluster 15.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
deep_learn	763	deep_learn	500
convolutional_neural_network	299	convolutional_neural_network	224
computer_vision	170	computer_vision	122
transformer	151	transformer	133
natural_language_process	139	natural_language_process	68
visual_question_answer	136	visual_question_answer	130
attention_mechanism	128	attention_mechanism	119
self_supervise_learn	108	self_supervise_learn	103
generative_adversarial_network	97	generative_adversarial_network	86

transfer_learn	95	transfer_learn	87
contrastive_learn	91	contrastive_learn	79
image_caption	89	image_caption	87
deep_neural_network	85	deep_neural_network	51
datum_augmentation	69	datum_augmentation	68
object_detection	64	object_detection	48
semantic_segmentation	59	semantic_segmentation	50

The results of this table could be interpreted as follows: “convolutional_neural_network” is the most frequently encountered “deep_learn” method in studies on “computer_vision”. An example of the work could be [7]. And “transformer” and “natural_language_process” are the most frequent “visual_question_answer” methods in “visual_question_answer” studies.

Table 4. INNER JOIN of top 20 terms from VOSviewer cluster 6 against top 20 terms from GSDMM cluster 24.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
datum_visualization	971	datum_visualization	421
visual_analytic	644	visual_analytic	105
datum_mine	134	datum_mine	83
interactive_visualization	83	interactive_visualization	41

Table 5. INNER JOIN of top 20 terms from VOSviewer cluster 8 against top 20 terms from GSDMM cluster 24.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
datum_analysis	111	datum_analysis	60
sentiment_analysis	67	sentiment_analysis	40
dashboard	50	dashboard	43
text_mine	36	text_mine	33

In this case, two clusters obtained from VOSviewer were compared with data from a GSDMM cluster. In both tables, the topic is close to “data analysis”. An example of the work could be [8].

Table 6. INNER JOIN of top 20 terms from VOSviewer cluster 1 against top 20 terms from GSDMM cluster 25.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
visualization	1061	visualization	292
information_visualization	283	information_visualization	146
virtual_reality	220	virtual_reality	188
augment_reality	180	augment_reality	118
human_computer_interaction	95	human_computer_interaction	71
eye_track	67	eye_track	50
model	58	model	33

evaluation	50	evaluation	31
immersive_analytic	50	immersive_analytic	46
perception	45	perception	34
mix_reality	43	mix_reality	34
design	39	design	35
user_experience	38	user_experience	33
storytell	33	storytell	31

If we exclude terms of the general type “visualization” and “information_visualization”, the most characteristic terms of this table are “virtual_reality” and “augmented_reality”. An example of the work could be [9].

Table 7. INNER JOIN of top 20 terms from VOSviewer cluster 3 against top 20 terms from GSDMM cluster 26.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
big_datum	182	big_datum	36
internet_of_thing	147	internet_of_thing	112
digital_twin	105	digital_twin	76
iot	72	iot	49
build_information_model	65	build_information_model	47
industry_4.0	60	industry_4.0	47
smart_city	45	smart_city	29
cloud_compute	39	cloud_compute	21
sensor	35	sensor	28
bim	34	bim	34

The term “iot” is an abbreviation for “internet_of_thing” they frequently occur in the records of the same cluster. This indicates the need to compile a dictionary of abbreviations, which, along with lemmatization, it is advisable to include in future works. An example of the work could be [10]. BIM (Building Information Model) is a rather specific term, so it has the same frequency of occurrence in the left and right columns.

Table 8. INNER JOIN of top 20 terms from VOSviewer cluster 6 against top 20 terms from GSDMM cluster 27.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
visual_analytic	644	visual_analytic	37
ccs_concept	152	ccs_concept	151
visualization_technique	44	visualization_technique	15
human_center_compute_visual_analytic	28	human_center_compute_visual_analytic	28
visualization_system_and_tool	23	visualization_system_and_tool	21

The term visual_analytic occurs frequently in the bibliometric records used in this paper, it occurs 644 times in the VOSviewer cluster and only 37 times in the specific cluster of GSDMM, the rest of the occurrences of this term are in other GSDMM clusters. An example of the work could be [11].

Table 9. INNER JOIN of top 20 terms from VOSviewer cluster 7 against top 20 terms from GSDMM cluster 28.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
particle_image_velocimetry	511	particle_image_velocimetry	508
flow_visualization	389	flow_visualization	374
piv	89	piv	87
computational_fluid_dynamic	84	computational_fluid_dynamic	78
cfD	67	cfD	62
heat_transfer	61	heat_transfer	58
two_phase_flow	57	two_phase_flow	57
flow_boil	50	flow_boil	49
numerical_simulation	43	numerical_simulation	33
turbulence	36	turbulence	35
proper_orthogonal_decomposition	34	proper_orthogonal_decomposition	34
3d_print	32	3d_print	23
large_eddy_simulation	32	large_eddy_simulation	31
flow_pattern	30	flow_pattern	30
fluid_structure_interaction	29	fluid_structure_interaction	28
aerodynamic	28	aerodynamic	27
flow_control	28	flow_control	28
pressure_drop	28	pressure_drop	26
heat_transfer_enhancement	24	heat_transfer_enhancement	24

In this case, the overlap of terms in the clusters is the highest of all 10 comparison tables. This is due to the fact that the cluster “particle_image_velocimetry” in the VOSviewer results is the most “freestanding”, i.e. its terms rarely co-occur with terms of other clusters. An example of an article reflecting the subject matter of this table is [12].

Table 10. INNER JOIN of top 20 terms from VOSviewer cluster 4 against top 20 terms from GSDMM cluster 29.

VOSviewer-AuKWs	N	GSDMM-AuKWs	N
machine_learn	599	machine_learn	308
neural_network	132	neural_network	54
cluster	130	cluster	112
dimensionality_reduction	126	dimensionality_reduction	122
classification	118	classification	67
feature_selection	35	feature_selection	34
random_forest	33	random_forest	32

The terms machine_learn and neural_network are often found in combination with other terms, so their frequency is greater in the left column than in the right column. It is interesting to note that among the specific machine learning methods, random_forest is found in the table, which is often used in practical applications [13].

Conclusions

A technique for visualizing the results of Author keywords clustering by the GSDMM algorithm is proposed, which facilitates the evaluation of the topics of records/documents in the cluster. It is based on comparing the occurrence of keywords of a given cluster and other clusters.

It is shown that the topics of clusters obtained with GSDMM algorithm can be corresponded with the results of clustering of Author keywords performed with the use of VOSviewer program.

The obtained results can be interpreted as a certain stability of the themes of clusters obtained by essentially different methods.

For further development of this study, the author considers it appropriate to:

1. to compile a thematic dictionary of abbreviations that allows unambiguous interpretation of terms such as IoT and the Internet of Things;
2. extend the obtained method of visualizing clustering results to other short texts — titles and abstracts;
3. analyze the impact of dictionary compilation on the clustering results of the GSDMM algorithm.

References

1. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction 2018. <https://doi.org/10.48550/ARXIV.1802.03426>.
2. Wang M, Vijayaraghavan A, Beck T, Posma JM. Vocabulary Matters: An Annotation Pipeline and Four Deep Learning Algorithms for Enzyme Named Entity Recognition. *J Proteome Res* 2024;23:1915–25. <https://doi.org/10.1021/acs.jproteome.3c00367>.
3. Yin J, Wang J. A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York New York USA: ACM; 2014, p. 233–42. <https://doi.org/10.1145/2623330.2623715>.
4. 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>.
5. Elshehaly M, Randell R, Brehmer M, McVey L, Alvarado N, Gale CP, et al. QualDash: Adaptable Generation of Visualisation Dashboards for Healthcare Quality Improvement. *IEEE Trans Visual Comput Graphics* 2021;27:689–99. <https://doi.org/10.1109/TVCG.2020.3030424>.
6. Ding X, Yang Z. Knowledge mapping of platform research: a visual analysis using VOSviewer and CiteSpace. *Electron Commer Res* 2022;22:787–809. <https://doi.org/10.1007/s10660-020-09410-7>.
7. Minaee S, Minaei M, Abdolrashidi A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors* 2021;21:3046. <https://doi.org/10.3390/s21093046>.
8. Camacho D, Panizo-LLedot Á, Bello-Orgaz G, Gonzalez-Pardo A, Cambria E. The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion* 2020;63:88–120. <https://doi.org/10.1016/j.inffus.2020.05.009>.
9. Yang Y, Dwyer T, Marriott K, Jenny B, Goodwin S. Tilt Map: Interactive Transitions Between Choropleth Map, Prism Map and Bar Chart in Immersive Environments. *IEEE Trans Visual Comput Graphics* 2021;27:4507–19. <https://doi.org/10.1109/TVCG.2020.3004137>.
10. Yu W, Dillon T, Mostafa F, Rahayu W, Liu Y. A Global Manufacturing Big Data Ecosystem for Fault Detection in Predictive Maintenance. *IEEE Trans Ind Inf* 2020;16:183–92. <https://doi.org/10.1109/TII.2019.2915846>.
11. Spinner T, Schlegel U, Schafer H, El-Assady M. explAiner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Trans Visual Comput Graphics* 2019:1–1. <https://doi.org/10.1109/TVCG.2019.2934629>.
12. Harte NC, Obrist D, Versluis M, Jebbink EG, Caversaccio M, Wimmer W, et al. Second order and transverse flow visualization through three-dimensional particle image velocimetry in millimetric ducts. *Experimental Thermal and Fluid Science* 2024;159:111296. <https://doi.org/10.1016/j.expthermflusci.2024.111296>.
13. Jiang F, Huo L, Chen D, Cao L, Zhao R, Li Y, et al. The controlling factors and prediction model of pore structure in global shale sediments based on random forest machine learning. *Earth-Science Reviews* 2023;241:104442. <https://doi.org/10.1016/j.earscirev.2023.104442>.

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.