UDC 004.912 DOI https://doi.org/10.35546/kntu2078-4481.2024.4.47

V. O. KHOLIEV
Postgraduate Student at the Department of Electronic Computers
Kharkiv National University of Radio Electronics
ORCID: 0000-0002-9148-1561

O. YU. BARKOVSKA
Associate Professor at the Department of Electronic Computers
Kharkiv National University of Radio Electronics
ORCID: 0000-0001-7496-4353

MODIFICATION OF THE METHOD OF LARGE TEXT SETS CLUSTERING

In this paper, a comparative analysis of common clustering methods such as k-means, Latent Dirichlet Distribution or LDA, Hierarchical Clustering Algorithm or HC, Density-based spatial clustering of applications with noise or DBSCAN, and Gaussian Mixture Model or GMM was conducted. The analysis was performed according to the selected criteria, such as scalability, computational complexity, presence (or absence) of a predefined number of clusters, and the evaluation approach (absolute with a clear relation to the cluster or relative using probabilities). According to the results, the DBSCAN method was chosen for further consideration due to a number of advantages, and the modification mod_DBSCAN was proposed, which reduces the number of potential calculations at each iteration, as a result, reduces computational complexity, and also increases system performance in conditions of limited resources. The modification consists of two changes: the vectorization stage, which is based on the annotation and keywords of the text specified by the author instead of the full text, and distance estimation for the so-called noisy points, which is performed in two steps. Popular datasets for the clustering task were analyzed. The proposed modification was tested on own Academ Lib Set dataset, formed on the basis of materials in the electronic catalog of the NURE scientific library. The analysis of the results showed an improvement in Precision by 5.6%, Recall by 12.5%, and F-score by 9.65%, which proves the effectiveness of the proposed modification. Further developments include testing combinations of methods and modules into larger functional blocks to identify and eliminate potential problems, as well as further optimization of such blocks. A separate work will investigate the approach to re-clustering after the dataset is updated. The quality of the new distribution is planned to be assessed based on the Rand index.

Key words: *clustering, clusterization, classification, dbscan, k-means, text processing, preprocessing, text, accuracy, cluster.*

В. О. ХОЛЄВ
аспірант кафедри електронних обчислювальних машин
Харківський національний університет радіоелектроніки
ORCID: 0000-0002-9148-1561

О. Ю. БАРКОВСЬКА
доцент кафедри електронних обчислювальних машин
Харківський національний університет радіоелектроніки
ORCID: 0000-0001-7496-4353

МОДИФІКАЦІЯ МЕТОДУ КЛАСТЕРИЗАЦІЇ ВЕЛИКИХ ТЕКСТОВИХ МАСИВІВ

У даній роботі було проведено порівняльний аналіз розповсюджених методів кластеризації, таких як k-means, Латентний розподіл Діріхле або LDA, Ієрархічний алгоритм кластеризації (IC), Density-based spatial clustering of applications with noise або DBSCAN, а також Модель суміші гаусіан (Gaussian Mixture Model або GMM). Аналіз проводився згідно з обраними критеріями, такими як масштабованість, обчислювальна складність, наявність (чи відсутність) умови попередньо визначеного числа кластерів, а також підхід до оцінювання (абсолютний з чітким відношення до кластеру або відносний з використання вірогідностей). Згідно з результатами, для подальшого розгляду був обраний метод DBSCAN через ряд переваг та зрештою була запропонована модифікація, яка зменшує кількість потенційних обчислень на кожній ітерації, як наслідок зменшує обчислювальну складність, що у свою чергу підвищує продуктивність системи, особливо в умовах обмежених ресурсів. Модифікація полягає у двох змінах: етапі векторизації, яка відбувається за анотацією та ключовими словами тексту, вказаними автором, замість повного тексту та оцінки відстані для так званих шумних точок, яка відбувається у два етапи. Запропоновану модифікацію методу DBSCAN було випробувано на власному датасеті Academ Lib Set, сформованому на основі матеріалів які знаходяться в електронному каталозі наукової бібліотеки ХНУРЕ. Аналіз результатів показав покращення результатів показників Precision на 5,6%, Recall на 12,5% та F-міри на 9,65%, що

доводить дієвість запропонованої модифікації. Подальші кроки передбачають випробування поєднань методів та модулів у більші функціональні блоки для виявлення та усунення потенційних проблем, а також подальша оптимізація таких блоків. Окремим роботою передбачається дослідження підходу до повторної кластеризації після оновлення набору даних з першочерговим розглядом підходів повторної кластеризації усіх документів (поточний метод) та такої для точок з найменшим силуетним коефіцієнтом. Оцінка якості формування нового розподілення планується на основі індексу Ранда.

Ключові слова: *кластеризація, класифікація, dbscan, k-means, обробка тексту, препроцесінг, текст, точність, кластер.*

Introduction

It is known that information has been accumulated since ancient times. The vast majority of information that was created was written, i.e., textual. And although since the mid-to-late twentieth century, with the development of technology, it has become possible to easily create, transmit, use and convert audio-visual information, huge amounts of already created textual information have remained at the disposal of humanity, and the textual type of information is actively used in all spheres of life. One of these areas is the academic and scientific sphere, where knowledge is still accumulated and transformed in textual form. This is due not only to the psychological and cultural traditions of mankind, but also to the fact that this form of information is the densest in terms of information load, unlike audio or visual formats.

Because of this, librarianship in general and the task of organizing and storing information in an organized manner in particular remains relevant, especially in the context of the overall increase in the volume of information generation with the emergence and development of the World Wide Web. In addition to librarianship, categorization is actively used in other areas that work with information (Figure 1).

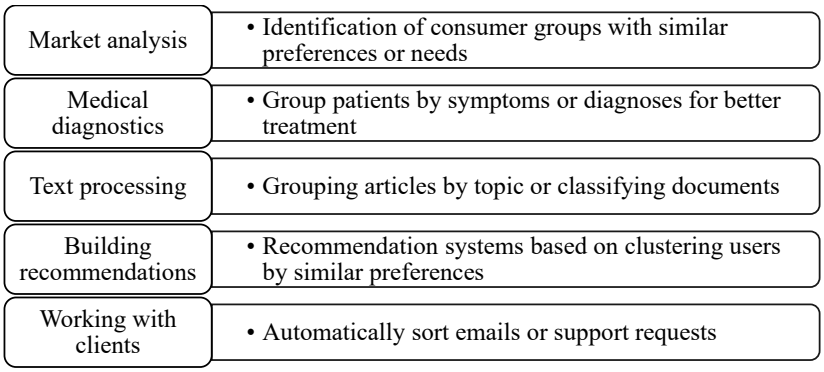


Fig. 1. Text categorization application areas

The organization process may include different steps depending on the goals and objectives, as well as the design of the structure, but it almost always includes a step of categorizing information, i.e., assigning it to a certain category.

While in the periods of analog librarianship, such a process was classification in nature – creating a list of classes and assigning information to one (or more) classes based on a number of criteria – in the digital era, with the development of computing power and an unprecedented increase in the amount of information that needs to be categorized, another type of categorization has become available on a practical level – clustering (Figure 2).

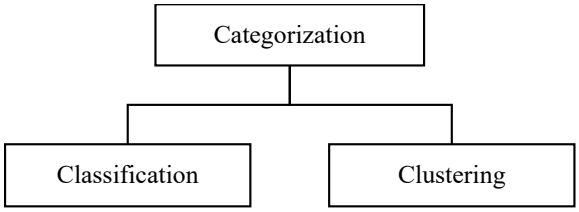


Fig. 2. General types of categorizations

During clustering, the entire set of information, or some part of it is analyzed, in the process of which, based on pre-formed or dynamically discovered criteria, the algorithm independently selects clusters and assigns information to them [1-2].

When it comes to categorizing information, criteria mean the properties of information units, or rather the difference in states of these properties.

While the properties of information units, such as subject matter or number of pages, are generic and inert in themselves, it is the difference in values or specific values of these properties that are the criteria for categorization, as they are comparable (e.g., what is the subject matter, what is the number of pages, is it greater than a given threshold, etc.).

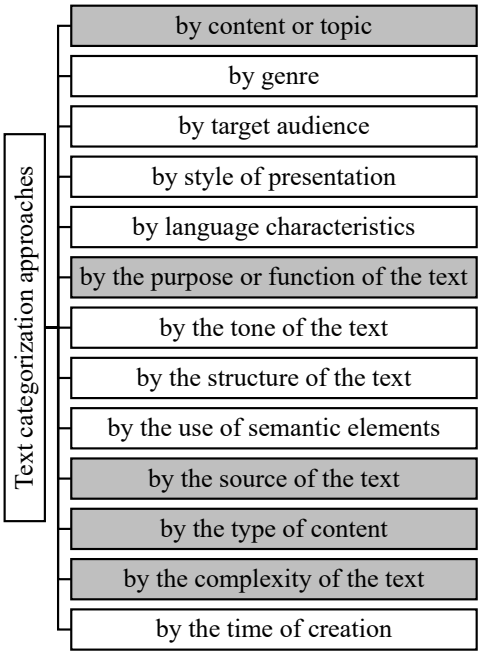


Fig. 3. Systematization of approaches to text categorization

In the context of information categorization, and more specifically, the clustering approach, a cluster is defined as a homogeneous group of objects that are very similar according to specified criteria and very dissimilar to other clusters of objects. A cluster has the following mathematical characteristics: center, radius, standard deviation, and cluster size. In turn, the center of the cluster is the average geometric location of the points in the space of variables, and the radius of the cluster is the maximum distance of the points from the center of the cluster.

Clusters can be overlapping. In this case, it is impossible to unambiguously assign an object to one of the two clusters using mathematical procedures. Such objects are called ambiguous, i.e., those that can be assigned to several clusters to the extent of similarity.

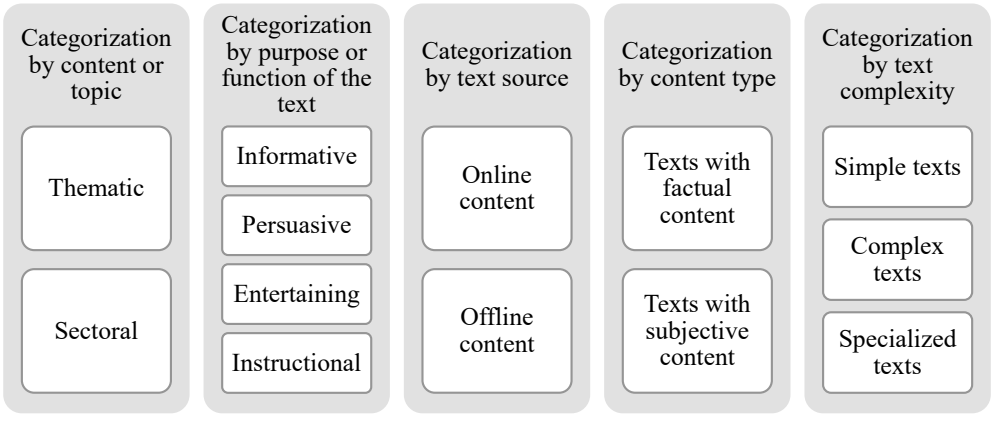


Fig. 4. Breakdown of sources by category type

The size of a cluster can be determined either by the cluster radius or by the standard deviation of the objects in the cluster. An object belongs to a cluster if the distance from the object to the center of the cluster is less than the cluster radius. If this condition is met for two or more clusters, the object is considered ambiguous.

Therefore, the condition for the emergence of a new cluster is the case when the distance of the object from the centers of all clusters is greater than their respective radii.

The result of clustering is a list of clusters and the units of information that belong to them. Another common way to visualize the results is to represent the objects as points and build a two-dimensional or three-dimensional graph – the so-called "cluster map".

This approach is useful when the categorization criteria are known in advance, but the final categories are not known, for example, when working with an unsorted data set. This is a suitable use case for a knowledge sharing system for young scientists, which was presented in [3] for the task of distributing the contingent of users according to the criterion of the direction of scientific interests, which are expressed in the topics and directions of their scientific works.

Related works

As was previously described, the clustering problem is not new, and despite the lack of practical implementation at the time, many solutions to the clustering problem have been proposed in the form of one or another approach. The following are popular and widespread clustering methods.

K-means

One of the most popular and simplest clustering algorithms. It divides the data into a predefined number of clusters, determining the centroid for each of them randomly or using some method (for example, k-means++). After the initial centroid calculation, each data point is assigned to the cluster whose centroid is closest to it. Next, new centroids are calculated for each cluster. Reassignment of clusters to data points and replacement of centroids is repeated until the centroids stop changing or until a specified number of iterations is reached. The advantages of k-means include simplicity and efficiency, but the algorithm requires a predefined number of clusters and is sensitive to outliers. In addition, it assumes that the clusters are spherical, which does not always correspond to the real-world data.

Density-based spatial clustering of applications with noise (DBSCAN)

The DBSCAN algorithm is based on the idea that a cluster in the data space is a continuous region of high point density separated from other similar clusters by continuous regions of low density [4]. To implement this, it uses two parameters – the maximum distance between points to be considered neighboring, and the number of points beyond which the density will be considered high and, accordingly, part of the cluster. To begin with, a random point in the set is selected and the number of neighboring points within the specified distance is calculated. If it is equal to or exceeds the minimum parameter, then all the found points (including the random one) are considered part of a newly defined cluster. After that, this step is recursively repeated for other points in the cluster until the number of neighboring points is less than the minimum. In this case, another random unprocessed point is selected, and the steps are repeated. The special feature of this algorithm is that it allows points that have not been assigned to any cluster, the so-called "noise points". The DBSCAN algorithm is well suited for datasets with unusual distribution shapes and also handles outliers well, but it is highly sensitive to the chosen parameter values.

Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation [5] is an unsupervised machine learning algorithm used in natural language processing to identify hidden topics present in a large corpus of documents. It works by assigning each document to a set of topics, and then uses a generative probabilistic model to determine the probability that a particular word in the document belongs to a particular topic. The algorithm uses two parameters – the number of topics and the distribution of words in each topic. The model assumes that there is a fixed set of topics (called the "preliminary") that are common to all documents, and for each document it looks for the distribution of these topics.

Hierarchical clustering (HC)

This method works as follows: first, each point is assigned to its own cluster. Then, using a specific proximity algorithm, the clusters closest to each other are determined and added to a new merged cluster. This is repeated until there is only one single remaining cluster that contains all the points. To determine the final distribution of clusters, a so-called dendrogram is constructed, which corresponds to the sequence of cluster merging. The final distribution is determined at the boundary between the two most distant mergers according to the constructed dendrogram. In this case, the cluster distance algorithm plays a crucial role.

Gaussian Mixture Model (GMM)

A GMM is a type of unsupervised learning algorithm, called so because it assumes that the data points to be clustered are not labeled with a value to be predicted. A GMM is usually expressed as a mixture of Gaussians, where each component represents a single variable. Each Gaussian is a probability density function that defines the probability that a data value falls within a certain distribution. The model assigns a probability to each cluster, which indicates the probability of a data point belonging to that cluster. GMM is capable of detecting clusters in data containing multiple overlapping distributions. For example, if a dataset contains data points that are grouped into two different categories, GMM can separate them into two separate clusters.

To analyze the above methods, it is necessary to determine the criteria for selecting the optimal one. In the framework of this work, the appropriate criteria are:

– scalability (the ability of the clustering algorithm to efficiently process large amounts of data or adapt to the growth of the input data size);

– computational complexity (describes the amount of time it takes for a clustering algorithm to perform its work with respect to a certain amount of data);
– the requirement to pre-determine the number of clusters;
– evaluation approach (absolute with a clear cluster association or relative using probabilities).

In accordance with the specified criteria, it is necessary to analyze the existing methods of text clustering and select the one or those that best meet them. The comparative analysis is presented in the table 1.

Table 1

Overview of the clustering methods characteristics

Method	Scalability	Comp. complexity	Pre-set number of clusters	Evaluation approach	Outstanding cons
K-means	good	*O(d*k*f)*, whered – number of docs in the corpus, k – number of clusters, f – number of features	yes	absolute	Sensitive to outliers; Spherical clusters shape
LDA	bad	O(d*n*t), where d – number of docs in the corpus, n – average number of words in documents, t – number of topics	yes	probabilistic	Describes each document as a set of topics;Predetermined set of topics
HC	bad	$O(d^2)$, where d – number of docs in the corpus	no	absolute	High computational complexity
DBSCAN	good	*O(d*log(d))*, where d – number of docs in the corpus	no	absolute	Sensitive to parameter values
GMM	bad	$O(d*g*f^2)$, where d – number of docs in the corpus,g – number of Gaussians,f – number of features	yes	probabilistic	Probabilistic approach to clustering;The need for a large number of sets for more accurate distribution.

Based on the results in the table and in accordance with the specified criteria, the DBSCAN method was chosen for analysis in this paper due to such properties as the absence of an initial condition on the number of clusters, the ability to form clusters of complex shapes and different sizes, and relatively lower computational complexity.

Similar studies have already been conducted [6-9], but none of them tested various clustering methods on a scientific and academic dataset. Therefore, testing and further modification of the method to take into account the potential features of such a dataset is a relevant task.

Aims and Tasks of the Work

The aim of this paper is to modify the clustering method of large text sets for the purpose of its further use in the text processing module of the knowledge exchange system for young scientists for the task of primary clustering of the user pool, as well as for the task of distributing further additions to existing or new clusters.

To achieve this goal, the following tasks must be accomplished:
– review datasets that will include scientific publications, articles and papers;
– conduct a comparative analysis of clustering methods according to the specified criteria;
– perform basic clustering of the selected dataset based using the DBSCAN method to determine the benchmark performance indicators (baseline);
– propose a modification of the DBSCAN method using the decomposition of input data taking into account the architecture of the computer system, which will reduce computational complexity and speed up the algorithm in conditions of limited system resources;
– analyze the results obtained.

Since this paper is the final one in the cycle of describing the methods used in the proposed system [3], the next steps include testing combinations of methods and modules into larger functional blocks to identify and eliminate potential problems, as well as further optimization of these blocks. A separate study is planned to investigate the approach to re-clustering after updating the dataset, with priority consideration of approaches to re-clustering all documents (the current method) and one for points with the lowest silhouette coefficient. The quality of the new distribution is planned to be assessed based on the Rand index.

Results and Discussion

Before modifying the clustering method, we derive a generalized algorithm [6-9]. The input is the document text, which is first of all subject to preprocessing, consisting of tokenization stages – splitting the text into tokens (usually words), cleaning the resulting set of tokens from noise, such as stop words, and further lemmatization or stemming. Optionally, a blacklist can be used at the noise removal stage, with words that should not be filtered out. After that, the processed token set is vectorized using one of the appropriate methods. This stage is also called feature extraction. The resulting vector represents a point in the multidimensional feature space, and the distance of the values of these features

is usually used to assign to clusters or to select a new one (depending on the algorithm). A diagram of such a generalized algorithm is shown in Figure 5.

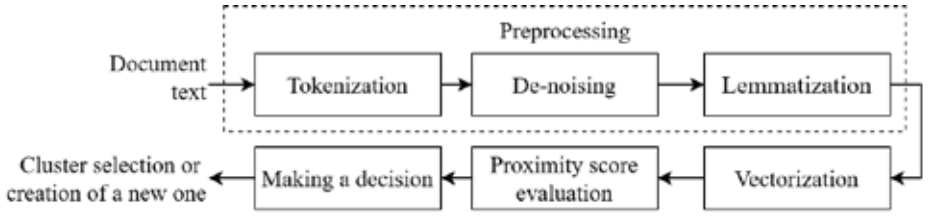
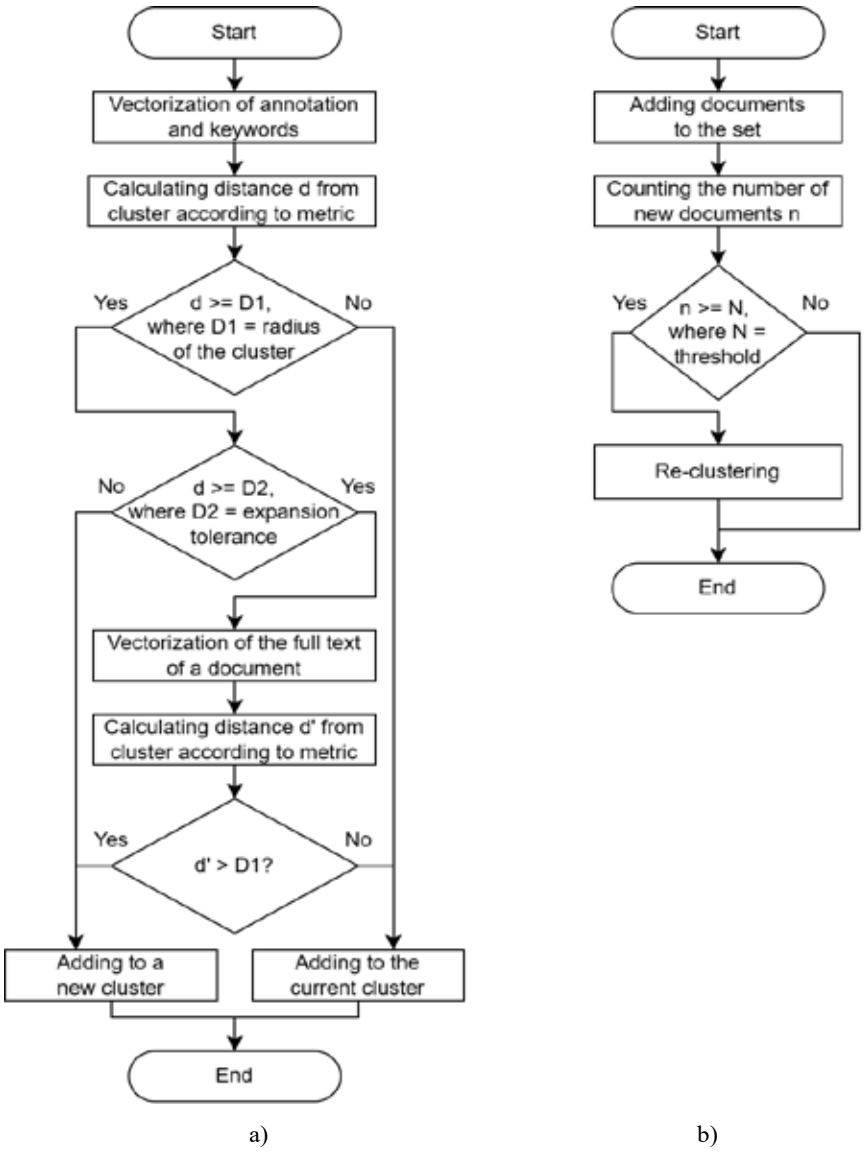


Fig. 5. Generalized clustering process

Based on the chosen DBSCAN clustering method, let's depict the proposed modification in the form of a flowchart (figure 6a). The proposed modification consists of two changes: the vectorization stage and distance estimation for the so-called noisy points. These modifications are aimed at reducing the number of potential computations, thereby reducing the computational complexity, which in turn will increase the system performance, especially in resource-constrained environments.



a) b)

Fig. 6. Flowcharts of the proposed solutions: a) two-stage clustering, б) condition of reclustering

Unlike the traditional vectorization stage, which usually vectorizes the full text of a document, the modified stage is applied only to the abstract and keywords provided by the author or authors of the document. This greatly reduces the computations that need to be performed at this stage for each document.

During the scanning of the epsilon radius, or the so-called threshold, for neighboring points, another, larger scanning radius is added – the tolerance of the threshold. Points that fall outside the traditional epsilon but within the tolerance range are calculated in an additional round of distance calculations. For this purpose, the full text of the document is vectorized and the epsilon distance is checked.

If a point based on a full-text vector falls within these boundaries, it becomes part of a cluster (figure 7).

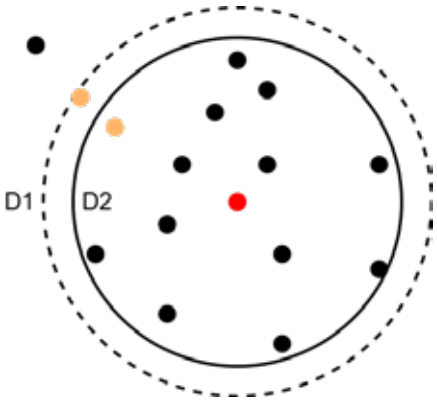


Fig. 7. Visualization of point configurations corresponding to the cases of the modified algorithm, where D2 is the threshold, D1 is the threshold tolerance

Since this algorithm is planned to be used in a knowledge sharing system for young scientists [3], it is necessary to provide a scenario of new raw data being added to the existing distribution. Since it is impractical to re-cluster for each new arrival, a naive algorithm is used that recalibrates the distribution after the arrival of a certain number of new objects (figure 6b).

Among the tasks of this study is to review the datasets to find a suitable one to test a DBSCAN clustering modification. Since obtaining real indicators of accuracy and F1-measure is possible only on real data with real topics that will ensure the operation of the information system for knowledge sharing of young scientists [3], the existing datasets were analyzed (table 2).

Table 2

Overview of datasets for the clustering task

Criterion	20 Newsgroups	Reuters-21578	AG News	Wikipedia Dump	Academ Lib Set
Amount of data	~20,000 documents	~21,578 documents	~120,000 documents	>10 mil documents	~5000 documents
Range of topics	News, technology, sports	Financial news, stock exchange, economy	News (world, sports, business, technology)	A wide range of topics: science, art, technology, etc.	Scientific and technical documents and publications
Language component	English	English	English	Multilingual	Ukrainian
Structure of documents	Text files grouped by topic	Annotated text files	Short text news	Heterogeneous (articles, tables, lists)	Heterogeneous (articles, tables, lists)
Annotation	Annotated	Annotated	Annotated	Partially annotated	Fully annotated
Rate of updates	Not updated	Not updated	Not updated	Annually	Bi-annually

The analysis of the comparison of existing datasets revealed the following drawback – the absence of a dataset with an exclusively scientific and technical focus. Therefore, own dataset was prepared on the basis of materials in the electronic catalog of the NURE scientific library [10], – Academ Lib Set.

Table 3

Experimental results

Method	Precision	Recall	F-міра
DBSCAN	0.783	0.696	0.715
mod_DBSCAN	0.827	0.783	0.784

Analyzing the values of the metrics – Precision (the proportion of correctly identified objects among all objects, showing the probability of no false positives), Recall (the proportion of correctly found objects among all real objects), F-measure (an assessment of the quality of text document clustering, combining Precision and Recall into one value) – we can see that the modified method shows the following positive increase in the values of the metrics: Precision by 5.6%, Recall by 12.5%, and F-measure by 9.65% (Figure 8).

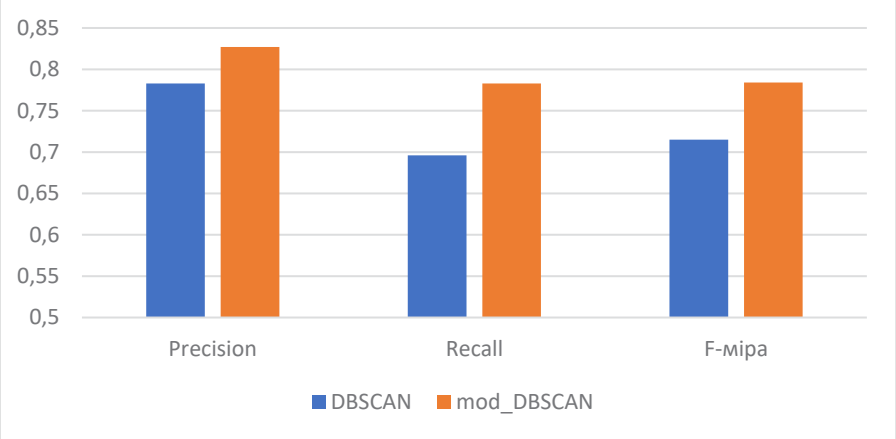


Fig. 8. Comparison of results for the basic and modified methods of clustering text documents DBSCAN and mod_DBSCAN

Conclusion

In this study, a comparative analysis of common clustering methods such as k-means, Latent Dirichlet Allocation or LDA, Hierarchical Clustering Algorithm (HC), Density-based spatial clustering of applications with noise or DBSCAN, and Gaussian Mixture Model (GMM) was conducted. The analysis was performed according to the specified criteria, such as scalability, computational complexity, presence (or absence) of a requirement for a predefined number of clusters, and the evaluation approach (absolute with a clear assignment to a cluster or relative using probabilities). According to the results, the DBSCAN method was chosen for further evaluation due to a number of advantages, and eventually a modification was proposed that reduces the number of potential calculations at each iteration, which consequently reduces the computational complexity, which in turn increases the system performance, especially in resource-constrained environments. The modification consists of two changes: the vectorization stage, which is based on the annotation and keywords of the text specified by the author instead of the full text, and distance estimation for the so-called noisy points, which is performed in two stages.

The proposed modification of the DBSCAN method was tested on own Academ Lib Set dataset, formed on the basis of materials in the electronic catalog of the NURE scientific library. The analysis of the results showed an improvement in Precision by 5.6%, Recall by 12.5% and F-measure by 9.65%, which proves the effectiveness of the proposed modification.

Further development involves testing combinations of methods and modules into larger functional blocks to identify and eliminate potential problems, as well as further optimization of such blocks. A separate work is planned to study the approach to re-clustering after updating the dataset, with priority consideration of approaches to re-clustering all documents (the current method) and one for points with the lowest silhouette coefficient. The quality of the new distribution is planned to be evaluated based on the Rand index.

Bibliography

1. Ahmed M. H., Tiun S., Omar N., Sani, N. S. Short Text Clustering Algorithms, Application and Challenges: A Survey. *Applied Sciences*. 2023. Vol. 13, No 1. P. 342. https://doi.org/10.3390/app13010342.

2. Dhar A., Mukherjee H., Dash N.S. та ін. Text categorization: past and present. *Artificial Intelligence Review*. 2021. Vol. 54. P 3007–3054. https://doi.org/10.1007/s10462-020-09919-1.

3. Барковська О., Холєв В., Пивоварова Д., Іващенко Г., Росінський Д. Система обміну знаннями молодих науковців із різних країн. *Сучасні інформаційні системи*. 2021. № 5(1). С. 69–74. https://doi.org/10.20998/2522-9052.2021.1.09.

4. Ester M., Kriegel H., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *AAAI Press* : In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), 1996. P. 226–231.

5. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 2003.Vol. 3, PP. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

6. Suyal H., Panwar A., Singh Negi A. Text Clustering Algorithms: A Review. *International Journal of Computer Applications*. 2014. T. 96, № 24. C. 36–40. URL: https://doi.org/10.5120/16946-7075.

7. Hotho A., Nürnberger A., Paaß G. A Brief Survey of Text Mining. *Journal for Language Technology and Computational Linguistics*. 2005. T. 20, № 1. C. 19–62. URL: https://doi.org/10.21248/jlcl.20.2005.68.

8. Zheng, Y., Cheng, X., Huang, R., Man, Y. A Comparative Study on Text Clustering Methods. *Springer, Berlin, Heidelberg* : In Advanced Data Mining and Applications. ADMA 2006, vol 4093. 2006. https://doi.org/10.1007/11811305_71.

9. Afzali M., Kumar S. Text Document Clustering: Issues and Challenges. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), м. Faridabad, 14–16 лют. 2019 р. 2019. URL: https://doi.org/10.1109/comitcon.2019.8862247.

10. Електронний каталог – Наукова бібліотека ХНУРЕ. Головна – Наукова бібліотека ХНУРЕ. URL: https://lib.nure.ua/el-katalog.

References

1. Ahmed, M. H., Tiun, S., Omar, N., & Sani, N. S. (2023). Short Text Clustering Algorithms, Application and Challenges: A Survey. Applied Sciences, 13(1), 342. https://doi.org/10.3390/app13010342

2. Dhar, A., Mukherjee, H., Dash, N.S. et al. Text categorization: past and present. Artif Intell Rev 54, 3007–3054 (2021). https://doi.org/10.1007/s10462-020-09919-1

3. Barkovska, O., Kholiev, V., Pyvovarova, D., Ivaschenko, G., & Rosinskiy, D. (2021). International system of knowledge exchange for young scientists. Advanced Information Systems, 5(1), 69–74. https://doi.org/10.20998/2522-9052.2021.1.09.

4. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 226–231.

5. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. Latent Dirichlet allocation // Journal of Machine Learning Research : journal / Lafferty, John. 2003. January (vol. 3, no. 4–5). P. pp. 993-1022. doi:10.1162/jmlr.2003.3.4-5.993

6. Suyal, Himanshu & Panwar, Amit & Negi, Ajit. (2014). Text Clustering Algorithms: A Review. International Journal of Computer Applications. 96. 36-40. doi:10.5120/16946-7075.

7. Hotho, Andreas & Nürnberger, Andreas & Paass, Gerhard. (2005). A Brief Survey of Text Mining. LDV Forum – GLDV Journal for Computational Linguistics and Language Technology. 20. 19-62. doi:10.21248/jlcl.20.2005.68.

8. Zheng, Y., Cheng, X., Huang, R., & Man, Y. (2006, August). A comparative study on text clustering methods. In International Conference on Advanced Data Mining and Applications (pp. 644-651). Berlin, Heidelberg: Springer Berlin Heidelberg.

9. Afzali, Maedeh & Kumar, Suresh. (2019). Text Document Clustering: Issues and Challenges. 263-268. 10.1109/COMITCon.2019.8862247.

10. Electronic catalog – NURE Scientific library. (n.d.). https://lib.nure.ua/el-katalog