A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms

Shapol M. Mohammed[1], Karwan Jacksi[2], Subhi R. M. Zeebaree[3]
[1]Department of Computer Engineering, Tishk International University (TIU), Erbil, Iraq
[2]Department of Computer Science, University of Zakho, Duhok, Iraq
[3]Duhok Polytechnic University, Duhok, Iraq

Article Info

Article history:

Received Nov 14, 2020
Revised Jan 19, 2021
Accepted Feb 15, 2021

Keywords:

DBSCAN
Density-based algorithm
DPC
GloVe word embedding
Similarity measurement
Evaluation measurement

ABSTRACT

Semantic similarity is the process of identifying relevant data semantically. The traditional way of identifying document similarity is by using synonymous keywords and syntactician. In comparison, semantic similarity is to find similar data using meaning of words and semantics. Clustering is a concept of grouping objects that have the same features and properties as a cluster and separate from those objects that have different features and properties. In semantic document clustering, documents are clustered using semantic similarity techniques with similarity measurements. One of the common techniques to cluster documents is the density-based clustering algorithms using the density of data points as a main strategic to measure the similarity between them. In this paper, a state-of-the-art survey is presented to analyze the density-based algorithms for clustering documents. Furthermore, the similarity and evaluation measures are investigated with the selected algorithms to grasp the common ones. The delivered review revealed that the most used density-based algorithms in document clustering are DBSCAN and DPC. The most effective similarity measurement has been used with density-based algorithms, specifically DBSCAN and DPC, is Cosine similarity with F-measure for performance and accuracy evaluation.

Corresponding Author:

Karwan Jacksi,
Department of Computer Science
University of Zakho
42002, Zakho, Duhok, Iraq
Email: Karwan.Jacksi@uoz.edu.krd

1. INTRODUCTION

Clustering is a technique used to group objects that have same properties and separate from other groups have different properties. In another mean maximize the similarity of data or objects inside one group and reduce the similarity between the objects of two groups [1-5]. Due to advances in a distributed system and parallel processing, the data size is increasing continuously [6-8]. This expansion is why data with huge size "big data" is available in different aspects [9, 10]. Analyzing big data require high-level knowledge and more accurate techniques [11, 12]. Clustering is using broadly in numerous fields, including data mining, information retrieval [13], knowledge discovery [14, 15] pattern recognition, and image segmentation. [16, 17] Mainly, document clustering is occupied a large area of research. In traditional document clustering, the keyword has an essential role in identifying the clusters. In contrast, in nowdays document clustering, semantic similarity has a significant role in identifying the relation between all objects in the same cluster [18-21].

In the literature clustering algorithms were classified into four major categories: the partitioning algorithm, density-based algorithm, grid-based algorithm, and hierarchical algorithm [22]. Performance, accuracy and speed of the density-based algorithms made the researchers widely depended on it in recent. A density-based algorithm depends on the idea of density of cluster's objects related to each other, spread in an arbitrary shape, and separated by low-density points of another cluster. It means that the primary notion of this approach is the density of objects. In the literature, there are many types of density-based algorithms, and the newest that has been used extensively is DBSCAN [22-24], and DPC [25-27].

Utilizing high dimensional or large-sized data as parameters for the clustering algorithms will result in low accuracy and unstable performance. Therefore, the best solution to this problem is using the word embedding representation to represent the corpus's data in a low dimentional vector space representation and to enhance the similarity of co-occurrences of word-to-word in the low dimensional vectors in a very efficient and expressive way. Recently, there are two widely known methods for creating a vector representation, which are word-to-vector (Word2Vec) and global vector (GloVe) [28, 29]. These two-methods stated that they are contesting to dedicate the literature and the most recently semantic similarity projects.

Pennington *et al.*, proved that GloVe is more efficient than Word2Vec, because the vector dimensionality of word2vec is low, and it cannot include all information from the corpus. In contrast, the GloVe is including the local and global information about the appeared words. The GloVe is unsupervised word embedding, introduced by Pennington *et al.*, from Stanford University for word representation. This algorithm uses the statistics of word-word co-occurrences in a corpus and is used for similarity and entity identification [29]. In this paper, we reviewed the most recent works and the more efficient techniques that have been developed and introduced.

2. LITERATURE REVIEW

As mentioned in the previous section, there are several types of document clustering techniques. The essential technique that has a significant role in document clustering is density-based clustering because it can discover clusters of arbitrary and different shapes [25]. Density-based clustering is a robust algorithm in analyzing specific data with a most remarkable performance [30], and it can provide adequate security in clustering data at various distributed datasets [31-33]. Document clustering has been used in many applications like text summarization [23, 24, 34], key phrase detection [35], and topic modeling [36, 37]. A literature review regarding the related works about GloVe word embedding and density-based clustering algorithms presented in the next two sections.

2.1. GloVe word embedding

The GloVe is a semantic vector space representation algorithm. It is representing documents based on word-word co-occurrences, and it is a count-based model. At the same time, the other word embeddings like word2vec is a predict based model. The GloVe is dependent on global matrix factorization and local context window. The GloVe performance is better than other word embeddings because it applies to non-zero elements and a subset of a corpus, not a whole corpus or a separate window of the significant corpus. GloVe outperforms other word embeddings in applications of word similarity and analogy and named entity recognition [38].

Document clustering has a significant role in natural language processing (NLP). The main stage in document clustering is the selection of convenient word representation of the dataset. In traditional word representation, TFIDF was used. TFIDF is a technique that calculates the frequency of each word in the data set and represents it in a one-dimension reducted vector to make a smaller feature space without measuring the relationships and dependency between words in the dataset. It causes ineffectiveness and hiding the semantics and syntactics of the words in the dataset. In recent years, many research types have been done on contextualized representation of data sets that have great affect on word vectors' semantic representation. Word2vec and GloVe are the two most populated word embeddings that use a co-occurrence frequency matrix of words to represent a one-dimension vector. Each line or row represents the vector representation of each word in the dataset. That can provide the best word embedding with rich features space [39]. Besides, Wrzalik and Krechel [40] used k-medoid to make the clusters balanced in having several members because k-medoid allows to inter-word similarity, word embeddings word2vec and GloVe have been used to represent the word vector of the datasets.

Tu *et al.*, [41] proposed an approach for domain-independent text segmentation for educational course content. First, optical character representation (OCR) has been used to convert speech to textual context, remove punctuation, and stop words for producing bag-of-words. Moreover, they used a GloVe algorithm to represent the texts of the document in a one-dimensional representation. Furthermore, the authors used topic modeling LDA to represent the shared topics of distributed documents. Then concatenate the output of the GloVe vector and LDA vector to a long vector representation. Also, they used Wasserstein distance to employ similarity between segments or long vectors to produce candidate blocks. Finally, they used Affinity propagation (AP) to block clustering the candidate blocks to final text segments. Naili *et al.,* [42] presented a comparative study of three methods LSA, Wort2Vec, and GloVe for modeling topic segmentation for the universal languages English and Arabic. They showed that Word2vec and GloVe were more effective than LSA. And Word2Vec was a little better than GloVe for low dimensional vector word representation.

Kamkarhaghighi and Makrehchi [43] proposed a content tree word embedding for document representation (CTWE) for enhancing the efficiency of the two famous word embedding (Word2Vec and GloVe). They aimed to improve F-score accuracy on the universal datasets (IMDB movie reviews and hate speech identification). Overall, the result of GloVe outperformed Word2Vec.

Bicalho *et al.* [36] presented a topic modeling on short text approach, they compared both (CoFE and DREx) algorithms for topic modeling. In term of DREx algorithm they compared the three algorithms of word vector representation (CBOW, GloVe and SG) their challenging work based on two significant points, which are co-occurrences of words and distributed word vector representation. The results showed that the GloVe has the seignifican effect on the rsults of topic modeling algorithm. At the same time, Wang *et al.*, [37] presented a two-stage of art topic modeling approach, which are the dirichlet multinomial mixture model (GSDMM) and latent feature latent dirichlet allocation (LDA). In the first stage, using (GSDMM) to measure word documents' probability and frequency to produce virtual documents, at the second stage, using a word embedding with LDA to extract each word's feature in virtual documents and produce the final tweet clusters. Also used both word embeddings (GloVe and Word2Vec) to represent word vectors of datasets then applied to clustering algorithm. The results showe that using GloVe make the topics more coherence than Word2Vec.

Wu and Li [44] presented a new approach for classifying documents, which is topic movers distance (TMD) instead of word movers distance (WMD). WMD has been used to clustering documents depending on word vector representation by GloVe algorithm for modeling a topic for each document. After that, TMD is used to classify documents depending on their topics, decreasing time complexity with WMD's with the same accuracy.

Saini *et al.*, [45] proposed a new automated multiobjective document clustering approach using a self-organized map (SOM) and TFIDF. Furthermore, used word embeddings like word2vec and GloVe to represent the data set in a vector space. The results showed that the proposed framework outperformed the single and multiobjective clustering algorithms and proposed approach is able to reach the global optimal solution for all the data sets, while other algorithms got stuck at local optima. The results clearly show that proposed framework is well suited for partitioning the data sets in an automated manner.

2.2. Density based clustering algorithms

Jang *et al.,* [47] proposed a density-based clustering algorithm for dialogue intention recognition system supported by word2vec word embedding to represent the enhancing similarity of corpus's data. Due to high dimensionality of word vectors size, the authors relied on T-distributed stochastic neighbor embedding (T-SNE) to convert the word vectors to a two-dimensional word vector. Hence, the size of the word vectors was reduced, it became very convenient for clustering. They used a density-based clustering algorithm to group data points. Besides, they utilized the DENCLUE algorithm over the DBSCAN because it was outperformed the DBSCAN algorithm in the clustering dialogue intention recognition specifically in term of performance evaluation.

Liao and Cheng [48] introduced a new approach for proving that semantic change is acquired from past to now by representing data in the google N-gram corpus. They used the word2vec word embedding algorithm to show the similarity relationship between all words in vector representation. They used density-based clustering specifically DBSCAN clustering algorithm to group similar words into clusters. Also, they visualized results of their approach to prove the semantic change.

Schmitt and Spinosa [49] presented sentiment analysis by using a convolutional neural network (CNN). The approach consisted of some layers and contained filters to fit the sentence and produce a related sentence to a matrix. Then, to be processed by word embedding (Word2Vec and GloVe) and presented the semantics of including words in the form of word vectors. However, the results of word2vec were chosen to be used for clustering due to the low dimensionality of GloVe, static and non-static models. Density-based clustering specifically DBSCAN was used to group semantic words in separate clusters and remove the outliers and noises. The experiments proved that removing outliers had a significant impact on the accuracy of the result.

Chen *et al.,* [50] suggested a framework of paragraph embedding for spoken document summarization. At first, they performed a comparison between paragraph embedding (SD and DBOW) and word embeddings (CBOW, SG, and GloVe). The result of paragraph embedding (SD and DBOW) outperformed word embeddings for summarizing documents. Four clustering algorithms were tested, including (VSM, MMR, DPC, and DPC_sum) to achieve good summarization results. The output of DPC and DPC_sum was impressive. The framework integrated the paragraph embeddings algorithm (SD and DBOW) with clustering algorithms (DPC and DPC_sum). The experimental result compared with existing and state of the art summarization frameworks, their result was excellent and satisfying.

Mary [51] proposed a dynamic density clustering algorithm, a developed version of DBSCAN and Chameleon algorithms. The author used Gaussian dimensionality to convert the real-time data in a dynamic dataset to a matrix of the two-dimensional size, to be convenient as a parameter of the clustering algorithms. The algorithm's result was proved in terms of accuracy and time by comparing it with the real DBSCAN algorithm.

Brown *et al.,* [11] proposed a new density-based algorithm to reduce the density algorithm's runtime. The concept of grid-based clustering and density-based clustering were integrated into one algorithm named a fast density-grid-based algorithm. The dataset converted to grid space each cell represented specific data then the density of grid cells calculated to produce clusters. The algorithm achieved its goal of reducing runtime without sacrificing its accuracy rate compared to the DBSCAN algorithm.

Cheng *et al.,* [52] proposed a new hierarchical clustering algorithm based on noise removal (HCBNR). The introduced algorithm used the concept of density-based clustering to remove the noise data from the corpus then partition the dataset into small clusters. Moreover, they used the hierarchical clustering algorithm to find the core clusters and then merge the small clusters and connect to their nearest core cluster. The algorithm's result compared to DBSCAN, DP, Chameleon, and CURE and outperformed those algorithms.

Harish *et al.,* [53] suggested an integration approach from two algorithms (SVC and FCM). The first one, which was a density-based clustering approach. SVC consisted of two concept Support Vector Clustering that can convert the data space from high dimensional space to low dimensional space. The second concept is density-based clustering, which can identify the cluster centers for each cluster and remove the outliers. The second part of the presented approach was fuzzy c-mean (FCM) algorithm, FCM would group classified documents based on the cluster centers identified by SVC.

Hou and Pelillo [54] analyzed the density peaks clustering (DPC). They presented the key factors that affected the algorithm's result, including manually selecting the number of clusters and radius of clusters. Also, the process of selecting the centers of the clusters. Those factors let the authors search and provide a solution to those issues. Hence, they developed the DPC algorithm to a level that can select the cluster centers accurately and provide a better result compared to traditional DPC, DBSCAN, and k-mean.

Liu *et al.,* [28] used the DPC for text document clustering in two steps. In the first step, they converted the text space into vector space using a supported vector machine (SVM) and cosine similarity to calculate the similarity among words. The second step used density peak clustering by calculating the local density and distance to determine the cluster center for each cluster and using a cutoff ratio to limit the cluster's circumstance. Applying the inside point to the cluster center as neighborhood and the outside point as noise and removed from the text space.

Patidar *et al.,* [55] presented an approach for activity detection from email metadata. Email metadata is including some information about users and activities they did. The approach consisted of two-stages. In the first stage converting the email metadata to a two-dimension matrix using SVD, then T-SNE have been used to reduce the dimentionality size of the outputted vectors, then they used DBSCAN algorithm to cluster the vectors. In the second stage, they performed a community detection algorithm on a network of clusters to merge the clusters that have the same properties and features to detect the last activities.

Lu and Zhu [56] proposed a framework based on density clustering to solve the problem of most existing clustering algorithms have been effected of having noises, different densities, different shapes and overlapping of clsueters. They used four modules to distribute the process of clustering. In the first module, the partition of the data space is divided into core and non-core points. In the second module, they initialized the core points to cluster centers. The third module ordering non-core points prioritized them to the next step, the fourth module classified non-core points to initial points. The proposed systems result is interesting compared with DBSCAN and DP results.

Yu *et al.,* [57], presented a developed version of DPC. They revised the real DPC by working on its limitation, which was the cutoff distance that must be pre-specified. They proposed a weighted local density sequence that could make the free propagation using k nearest neighborhood to detect neighborhood sequence. The developed algorithm's accuracy and performance were proved by comparing it with existing DPC modified algorithms.

Jiang *et al.,* [27] proposed a new density peak clustering technique using the gravitation density-based clustering (GDPC). According to the authors, there is a limitation on the DPC algorithm, including detecting anomalies. GDPC used gravitation theory based on natural gravitation force to detecting cluster anomalies with varying sizes and numbers of clusters. GDPC has a better performance compared to k-means, AP, and DPC. Thus, the original DPC has a problem of processing clustering over high dimensional data because in high dimensional data there are more interactive data with varying densities in different places.

Du *et al.,* [58] proposed an aggregated algorithm to solve the limataions of original DPC of does not working well on high dimentional data and ignoring local structure of data, which was density peak clustering (DPC-KNN-PCA) based on k-nearest neighbors (KNN), and principal component analysis (PCA). They used PCA to reduce high dimensional data to low dimensional data. Also, the KNN has been used to identify the cluster centers supporting DPC. The result of this approach was fessible and effective compared to k-means and spectral clustering (SC) algorithm in term of accuracy.

Wu *et al.,* [59] proposed a new algorithm based on the density of k-means, which was cluster center initialization based on density peak (CCIDP). They used the concept of density peak to initialize the cluster center of k-means algorithm, because k-means algorithm has a problem of hard identifying cluster centers. In the beginning, they used VSM to represent the data space as vector representation. Also, for vectors item data representation, they used TF-IDF to represent the item feature of each data in the vector. Moreover, for identifying the distance between the points, the authors used cosine similarity to calculate the similarity between points, then convert the measurements to distance measurement between data points, then performed the CCIDP algorithm on the vectors. The results showed that it is better and decreases the iteration number compared with existing algorithms.

3. DISCUSSION

A survy have been done on a collection of three types of researches: the first collection includes those researches that used GloVe word embedding algorithm in their approaches as show in Table 1. The second collection contained those approaches that used density-based clustering algorithms or the concept of density-based clustering in their approaches as shown in Table 2. The final collection consists of those approached that used the concept of word embedding or matrix representation of datasets as shown in Table 3. Besides, the ratio of each used algorithm in this survy has been calculated and presented in Figure 1. The most used algorithm in this survy is DBSCAN clustering algorithm with 31% of the survy researches based on it. Moreover, it is clear that 26% of the researches in this survy used the density peak clustering algorithm (DPC) in their approaches. The researchers' fewer utilized algorithms are AP and Chameleon, DENCLUE and SVC were used with 3% of researches. Also, 10% of the researchers were used K-mean. Hierarchical and KNN have been utilized by 8% of the researchers, while the Grid-based algorithm was used by 5% of literature in this survy.

Figure 2 presents the similarity measurements used by GloVe, DBSCAN, and DPC algorithms through this survy. The cosine similarity is widely used by the three algorithms. The GloVe and DBSCAN algorithms have peaked through using cosine similarity. However, with the DPC algorithm, the Euclidian distance has been used more than other algorithms and is not used by the DBSCAN. In contrast, the TF-IDF is not utilized by the DPC algorithm, but the TD-IDF is the second widely used similarity technique by the GloVe algorithm. In comparison, the Jaccard similarity has been used with the same level by the three algorithms.

In general, most of researches have been collected from the literature review related to word embeddings and density-based algorithms. Through the tables and figures, the following two points have been observed, i) global word embedding for document clustering is used in a smaller number of researches with density-based algorithms, and ii) density-based algorithms or the concept of density-based algorithms were used widely in document clustering, document summarization, and document topic modeling.

Another survey has been done on evaluation measurements that have been used by the three algorithms (GloVe, DBSCAN, and DPC). Evaluation measurement is a standard metric using for evaluating the accuracy and performance of clustering algorithms. The bar chart Figure 3 has been created by measuring each evaluation metric's frequency occuring from the tables. F-measure or F-score has a peak level of the three algorithms. In contrast, the entropy and (Dun and building index) have no level with the DPC algorithm.

For the GloVe word embedding, all the evaluation metrics have been used. F-measure has a peak level. Besides, the entropy metric has the second level. In comparison, all remaining metrics have the same level of usage with GloVe algorithm. DBSCAN algorithm has a peak level of using F-measure for evaluating its results. While entropy and (Dun and Buldin index) have the same and lowest level. However, recall and precision measurements have the same level and significantly fewer levels than F-measure.

In our survey, DPC has no level of using entropy and (Dun and Buldin). In contrast, the F-measure has a very long peak level than all other evaluation metrics frequencies. However, recall and precision have 1/3 level of F-measure in evaluating the result of DPC. Overall, the F-measure is the most widely used metric for evaluating the result of all three algorithms.

Table 1. GloVe word embedding for word representation

Article	Clustering algorithm	Similarity measure	Evaluation measure	Dataset
[41]	Unsupervised clustering Affinity Propagation (AP)	Wasserstein distance similarity		Online education course and Choi dataset
[36]		Cosine similarity	F-score	TMN, NBA, Politics, 20Nshort, Sanders, Snippets, and CLEF
[37]	A hierarchical agglomerative clustering algorithm	Cophenetic Correlation Coefficient (CPCC)	Normalized Mutual Information (NMI) and topic coherence	Story detection corpus and a large-scale event detection corpus covering over 500 events
[45]	K-mean	*Tf-IDF similarity*	*Dunn Index* and *Davies-Bouldin (DB) Index*	NIPS 2015, AAAI 2013, Webb
[39]	Own algorithm	Cosine similarity	Entropy	SQuAD 1.1, Yahoo Answers, REUTERS and FakeNewsAMT
[40]	*K*-medoid	Tf-IDF similarity	Sharp and fuzzy	Amazon, Classic, Reuters, 20News, and WIKI
[43]		Correlation	F-score	IMDB Movie Reviews and Hate Speech Identi_cation
[46]	K-mean	Cosine similarity, Euclidian distance, and Jaccard		news corpus and a national corpus of the Russian language
[38]		Cosine similarity	Entropy evaluation measure and all evaluation metrics, except for the CoNLL.	
[44]	KNN	Topic Mover's Distance (TMD)		Menu recipe, epic recipe twitter, bbc sport, bbc, and reuters

Table 2. Density-based algorithms

Article	Clustering algorithm	Similarity measure	Evaluation measure	Dataset
[51]	Dynamic clustering algorithm based on DBSCAN and chameleon	Generalized Dunn Index (GDI), Davies-Bouldin index (DB)		synthetic dataset and real data sets from UCI Data repository
[11]	Fast Density-Grid Clustering Algorithm		Result in accuracy and runtime comparison with DBSCAN according to its datasets.	Julia Handl's website datasets and UCI Machine Learning Repository datasets
[52]	Hierarchical clustering based on noise removal (HCBNR)		ACC and NMI for performance evaluation	Synthetic dataset and five datasets from UCI
[53]	C-mean		Precision, Recall, and F -measure	20-Newsgroup of the popular standard dataset for text categorization
[54]	Density peak clustering algorithm (DPC)	Jaccard index	F-score	Aggregation, R15, D31, Jain, Thyroid, Wdbc, Iris, Breast, and four datasets of non-2D datasets of UCI repository.
[28]	Density peak algorithm	cosine similarity	Precision, Recall and F-score	Reuters-21578
[55]	Two-stage algorithm SVD followed by DBSCAN	Jaccard and TFIDF	Precision, Recall, and F-Measure	Ground-truth dataset
[56]	The density clustering framework used density and partitioning algorithms together.		Compare results with DBSCAN and Dp algorithm	synthetic datasets (Compound, D31, DS3, DS4 and Flame) and real-world datasets (Cancer, Control, diabetes, iris, seeds, sonar, and wine)
[31]	Distributed density-based algorithm (k-nearest neighbors)	Own Distance measure (Linkage, Closeness, Sharing, and Connectivity)		Synthetic dataset
[57]	weighted local density sequence and two-stage assignment strategies, called DPCSA	Euclidean distance	ACC, AMI, ARI, and efficiency	artificial and real-world datasets and the well-known Olivetti face dataset Sharing,
[27]	Gravitation Density Peak Clustering (GDPC)		F- measure and comparative result with DPC, AP, and K-means	Four UCI datasets (Iris, Seeds, Wine, Glass) and Three synthetic datasets (Flame, Aggregation, and Spiral)

Table 2. Density-based algorithms (*continued*)

Article	Clustering algorithm	Similarity measure	Evaluation measure	Dataset
[60]	K-means, DBSCAN, HCA and MDBCA		Using WEKA to compare the performance of the four algorithms.	Abalone, Bankdata, Router, SMS and Webtk
[25]	Modified DBSCAN and gride based algorithm.		Comparing the performance of the algorithm with MCDAStream, stream and CluStream.	*Synthetic dataset* NMG (Noisy Mixture of Gaussians) and UCI dataset NID (Network Intrusion Detection)
[58]	DPC-KNN-PCA	Euclidean distance.	Performance compared to classical methods (k-means algorithm and spectral clustering algorithm.	Synthetic dataset (R15, Aggregation, Flame, and D) and UCI datasets (Iris, LED digits, Seeds, Heart, Pen-based digits, Waveform, and Sonar).
[26]	W-DBSCAN		Purity comparison with FCM and k-means	Visual datasets (spiral, flame, and aggregation) and UCI datasets (Iris, Tae, Cmc, and Seeds)

Table 3. Density-based algorithms, specifically DBSCAN and DPC with matrix representation of word documents

Article	Clustering algorithm	Word embedding	Similarity measure	Evaluation measure	Dataset
[47]	DENCLUE	Word2Vec + T-SNE		Precision, recall, f-score, the success rate, and separation rate.	raw corpus and an annotated dialogue corpus
[48]	DBSCAN	Word2Vec	Cosine similarity	Comparing and visualizing the past and new results	N-gram
[49]	DBSCAN	CNN + (Word2Vec and GloVe)	Cosine similarity	Comparing the result of a static and non-static model using two embeddings (word2vec and GloVe) over three different datasets.	Movie reviews (MR), subjectivity (Subj) and IMBD
[50]	DPC and DPC-sum	Distributed memory (SD) and distributed Bag-Of-words (DBOW)		They compared the result of (Sd and DBOW) with (CBOW, SG, and GloVe). Furthermore, comparing the result of (DPC and DPC-sum) with (VSM and MMR).at the same time comparing the overall result with the existing and state of the art algorithms result.	MATBN broadcast news
[22]	DBSCAN	WordNet + TF-IDF	Cosine similarity	F-measure and entropy	ACM and PubMed
[30]	DBSCAN		Cosine similarity		Twitter and Google map
[23]	DBSCAN			Recall, precision, and f-measure	News collection dataset
[24]	DBSCAN + HMM		wordNet, Mapper, and Reducer for semantic similarity	Recall, precision, and f-measure	British Medical Journal (BMJ)
[35]	DPC	DGM	Cosine similarity	Recall, precision, and f-measure	UCST-News and HULTH2003
[34]	DPC	Word/paragraph embedding		F-measure (ROUGE-1, ROUGE-2, and ROUGE-SU)	DUC2003 and DUC2004
[29]	DPC	BOW	Cosine similarity	F-measure (ROUGE-1 and ROUGE-2)	TAC 2008

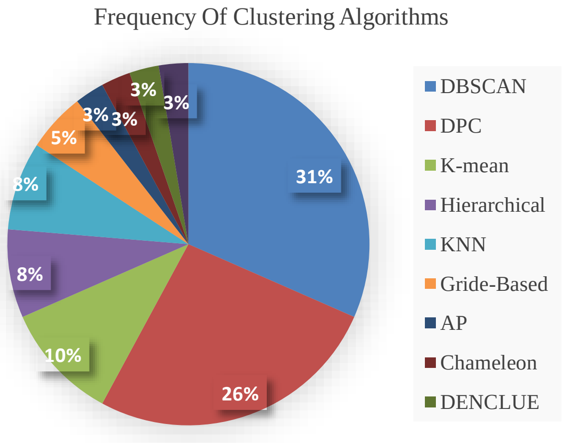
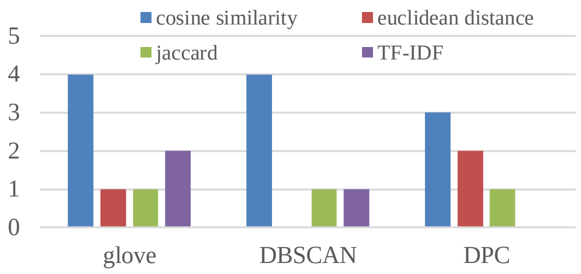


Figure 1. Describe the used ratio of each density algorithm in this survey



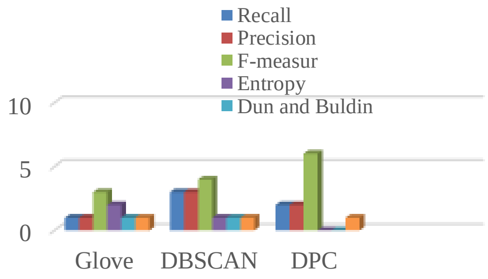


Figure 2. Similarity measurements according to the GloVe and density-based algorithms

Figure 3. Evaluation measurements according to the GloVe and density algorithms

4. CONCLUSION

In this survy a collection of researches from various databases between 2014 to 2020 have been gathered those were used the GloVe word embedding or concept of matrix data representation and density-based or concept of density-based clustering algorithms. This survey was conducted based on thirty-six papers to know which types of the density-based algorithms are more used specifically for document clustering with GloVe word embedding. Also, to know which types of similarity measurements and evaluation metrics were used by density-based algorithms and GloVe word embedding. In our experiments and based on the survey, GloVe word embedding is very little have been used with density-based algorithms. The two types of density-based algorithms that more used in our survey are DBSCAN and DPC. Also, in our survey, cosine similarity and F-measure are two metrics that are most used for similarity measurement and evaluation of the performance and accuracy of the two density-based algorithms.

5. FUTURE WORK

In future work, we will use global word embedding (GloVe) with DBSCAN clustering algorithms for developing semantic document clustering. Also, we will use cosine similarity to measure the similarity between words and F-measure to evaluate the model's accuracy and performance on the two datasets (Wikipedia and IMDB).

REFERENCES

[1] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[2] K. Jacksi and S. Badiozamany, "General method for data indexing using clustering methods," *International Journal of Scientific and Engineering Research (IJSER)*, vol. 6, no. 3, pp. 641-644, Mar. 2015.

[3] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and S. R. M. Zeebaree, "Combination of k-means clustering with genetic algorithm: A review," *International Journal of Applied Engineering Research*, vol. 12, no. 24, pp. 14238-14245, 2017.

[4] S. R. Zeebaree, K. F. Jacksi, and R. R. Zebari, "Impact analysis of SYN flood DDOS attack on HAPROXY and NLB cluster-base web servers," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 19, no. 1, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp%p.

[5] S. R. Zeebaree, R. R. Zebari, and K. Jacksi, "Performance analysis of IIS10. 0 and apache2 cluster-based web servers under syn ddos attack," *Test Engineering and Management*, vol. 83, no. 1, pp. 5854-5863, 2020.

[6] L. M. Haji, S. Zeebaree, O. M. Ahmed, A. B. Sallow, K. Jacksi, and R. R. Zeabri, "Dynamic resource allocation for distributed systems and cloud computing," *TEST Eng. Manag*, vol. 83, pp. 22417-22426, 2020.

[7] Z. N. Rashid, S. R. Zebari, K. H. Sharif, and K. Jacksi, "Distributed cloud computing and distributed parallel computing: a review," 2*018 International Conference on Advanced Science and Engineering (ICOASE)*, 2018, doi: 10.1109/ICOASE.2018.8548937.

[8] S. R. Zeebaree, H. M. Shukur, L. M. Haji, R. R. Zebari, K. Jacksi, and S. M. Abas, “Characteristics and analysis of hadoop distributed systems," *Technology Reports of Kansai University*, vol. 62, no. 4, pp. 1555-1564, 2020.

[9] K. Jacksi and S. M. Abass, "Development history of the world wide web," I*nternational Journal of Scientific & Technology Research*, vol. 8, no. 9, pp. 75-79, 2019.

[10] V. R. Benjamins, "Big data: from hype to reality?," i*n Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14),* Thessaloniki, Greece, Jun. 2014, pp. 1-2, doi:10.1145/2611040.2611042.

[11] D. Brown, A. Japa, and Y. Shi, "A Fast density-grid based clustering method," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA*, Jan. 2019, pp. 0048-0054, doi:10.1109/CCWC.2019.8666548.

[12] K. Jacksi, "Toward the semantic web and linked data exploration," *2019 4th Scientific International Conference Najaf (SICN)*, 2019, doi:10.1109/SICN47020.2019.9019361.

[13] K. Jacksi, S. R. M. Zeebaree, and N. Dimililer, "LOD explorer: presenting the web of data," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 1, 2018, doi:10.14569/IJACSA.2018.090107.

[14] K. Jacksi, N. Dimililer, and S. R. M. Zeebaree, "A Survey of Exploratory Search Systems Based on LOD Resources," in *Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015 11-13 August, 2015 Istanbul, Turkey,* 2015, pp. 501-509.

[15] K. Jacksi, N. Dimililer, and S. R. Zeebaree, "State of the art exploration systems for linked data: a review," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 11, pp. 155-164, 2016, doi:10.14569/IJACSA.2016.071120.

[16] N. D. Karwan Jacksi,Subhi R. M. Zeebaree, "AN improved approach for information retrieval with semantic-web crawling," PhD. Thesis. 2016.

[17] K. Jacksi, S. Zeebaree, and N. Dimililer, "Design and implementation of LOD explorer: a LOD exploration and visualization model," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 01-09, 2020, doi:10.38094/jastt1214.

[18] N. Y. Saiyad, H. B. Prajapati, and V. K. Dabhi, "A survey of document clustering using semantic approach," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 3-5 March 2016, doi:10.1109/ICEEOT.2016.7755154.

[19] N. M. Salih and K. Jacksi, "State of the art document clustering algorithms based on semantic similarity," *Jurnal Informatika*, vol. 14, no. 2, pp. 58-75, 2020, doi:10.26555/jifo.v14i2.a17513.

[20] R. Ibrahim, S. Zeebaree, and K. Jacksi, "Survey on semantic similarity based on document clustering," A*dv. Sci. Technol. Eng. Syst. J.*, vol. 4, no. 5, pp. 115-122, 2019, doi:10.25046/aj040515.

[21] R. Ibrahim, S. R. M. Zeebaree, and K. Jacksi, "Semantic similarity for document clustering using TFIDF and K-mean," Master’s Thesis, University of Zakho, Zakho, 2020.

[22] G. Veena and N. K. Lekha, "A concept based clustering model for document similarity," in 2*014 International Conference on Data Science & Engineering (ICDSE), Kochi, India, Aug. 2014,* pp. 118-123, doi:10.1109/ICDSE.2014.6974622.

[23] R. Reztaputra and M. L. Khodra, "Sentence structure-based summarization for Indonesian news articles," in 2*017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, Denpasar, Aug. 2017, pp. 1-6, doi:10.1109/ICAICTA.2017.8090983.

[24] K. T. Belerao and S. B. Chaudhari, "Summarization using mapreduce framework based big data and hybrid algorithm (HMM and DBSCAN)," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, Sep. 2017, pp. 377–380, doi:10.1109/ICPCSI.2017.8392320.

[25] K. Shyam Sunder Reddy and C. Shoba Bindu, "StreamSW: A density-based approach for clustering data streams over sliding windows," *Measurement*, vol. 144, pp. 14-19, Oct. 2019, doi:10.1016/j.measurement.2018.11.041.

[26] M. Huang, Y. Yan, L. Xu, and L. Ye, “Using warshall to solve the density-linked density clustering algorithm,” *American Journal of Applied Mathematics*, vol. 8, no. 1, pp. 11-16, February 2020, doi:10.11648/j.ajam.20200801.12.

[27] J. Jiang, D. Hao, Y. Chen, M. Parmar, and K. Li, "GDPC: gravitation-based density peaks clustering algorithm," P*hysica A: Statistical Mechanics and its Applications*, vol. 502, pp. 345-355, Jul. 2018, doi:10.1016/j.physa.2018.02.084.

[28] P. Liu, Y. Liu, X. Hou, Q. Li, and Z. Zhu, "A text clustering algorithm based on find of density peaks," in *2015 7th International Conference on Information Technology in Medicine and Education (ITME)*, Huangshan, China, Nov. 2015, pp. 348–352, doi:10.1109/ITME.2015.103.

[29] W. Guohua and G. Yutian, "Using density peaks sentence clustering for update summary generation," in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Vancouver, BC, Canada, May 2016, pp. 1-5, doi:10.1109/CCECE.2016.7726719.

[30] N. Pandey, "Density based clustering for Cricket World Cup tweets using Cosine similarity and time parameter," *2015 Annual IEEE India Conference (INDICON)*, New Delhi, India, 2015, pp. 1-6, doi:10.1109/INDICON.2015.7443520.

[31] A. Salim, "Density based clustering algorithm for distributed datasets using mutual k-nearest neighbors," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 3, 2019, doi:10.14569/IJACSA.2019.0100380.

[32] M. A. Sadeeq, S. R. Zeebaree, R. Qashi, S. H. Ahmed, and K. Jacksi, "Internet of things security: a survey," vol. 88, pp. 162-166, 2018, doi:10.1016/j.jnca.2017.04.002.

[33] S. R. Zeebaree, R. R. Zebari, K. Jacksi, and D. A. Hasan, "Security approaches for integrated enterprise systems performance: a review," *International Journal of Scientific & Technology Research*, vol. 8, no. 12, pp. 2485-2489, 2019.

[34] Baoyan Wang, Jian Zhang, Fanggui Ding, and Yuexian Zou, "Multi-document news summerization usning paragraph embedding and density peaks clustering," *2017 International Conference on Asian Language Processing (IALP)*, 5-7 Dec. 2017, doi:10.1109/IALP.2017.8300593.

[35] M. Alfarra, A. M. Alfarra and A. Salahedden, "Graph-based Density Peaks Ranking Approach for Extracting KeyPhrases (GDREK)," *2019 IEEE 7th Palestinian International Conference on Electrical and Computer Engineering (PICECE)*, Gaza, Palestine, 2019, pp. 1-6, doi: 10.1109/PICECE.2019.8747175

[36] P. Bicalho, M. Pita, G. Pedrosa, A. Lacerda, and G. L. Pappa, "A general framework to expand short text for topic modeling," *Information Sciences*, vol. 393, pp. 66-81, Jul. 2017, doi:10.1016/j.ins.2017.02.007.

[37] B. Wang, M. Liakata, A. Zubiaga, and R. Procter, "A hierarchical topic modelling approach for tweet clustering," *International Conference on Social Informatics*, 2017, doi:10.1007/978-3-319-67256-4_30.

[38] J. Pennington, R. Socher, and C. Manning, "GloVe: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532-1543, doi:10.3115/v1/D14-1162.

[39] J. Park, C. Park, J. Kim, M. Cho, and S. Park, "ADC: Advanced document clustering using contextualized representations," Expert Systems with Applications, vol. 137, pp. 157–166, Dec. 2019, doi:10.1016/j.eswa.2019.06.068.

[40] M. Wrzalik and D. Krechel, "Balanced word clusters for interpretable document representation," in 2*019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Sydney, Australia, Sep. 2019, pp. 103-109, doi:10.1109/ICDARW.2019.40089.

[41] Y. Tu, Y. Xiong, W. Chen, and C. Brinton, "A domain-independent text segmentation method for educational course content," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, Singapore, Singapore, Nov. 2018, pp. 320-327, doi:10.1109/ICDMW.2018.00053.

[42] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, "Comparative study of word embedding methods in topic segmentation," *Procedia Computer Science*, vol. 112, pp. 340-349, 2017, doi:10.1016/j.procs.2017.08.009.

[43] M. Kamkarhaghighi and M. Makrehchi, "Content tree word embedding for document representation," *Expert Systems with Applications*, vol. 90, pp. 241-249, Dec. 2017, doi:10.1016/j.eswa.2017.08.021.

[44] X. Wu and H. Li, "Topic mover's distance based document classification," *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, 27-30 Oct. 2017, doi:10.1109/ICCT.2017.8359979.

[45] N. Saini, S. Saha, and P. Bhattacharyya, "Automatic scientific document clustering using self-organized multi-objective differential evolution," *Cogn Comput*, vol. 11, no. 2, pp. 271-293, Apr. 2019, doi:10.1007/s12559-018-9611-8.

[46] R. Mussabayev, B. Kassymzhanov, A. Mukashev, V. Ibrayeva, and A. Merkebayev, "Creation of necessary technical and expert- analytical conditions for development of the information system of evaluating open text information sources' influence on society," in *2019 15th International Asian School-Seminar Optimization Problems of Complex Systems (OPCS), Novosibirsk, Russia*, Aug. 2019, pp. 104-109, doi:10.1109/OPCS.2019.8880193.

[47] J. Jang, Y. Lee, S. Lee, D. Shin, D. Kim, and H. Rim, "A novel density-based clustering method using word embedding features for dialogue intention recognition," *Cluster Comput*, vol. 19, no. 4, pp. 2315-2326, Dec. 2016, doi:10.1007/s10586-016-0649-7.

[48] Liao X., Cheng G., Analysing the Semantic Change Based on Word Embedding. In: Lin CY., Xue N., Zhao D., Huang X., Feng Y. (eds) Natural Language Understanding and Intelligent Applications. ICCPOL 2016, NLPCC 2016. Lecture Notes in Computer Science, vol 10102, 2016, Springer, Cham, doi:10.1007/978-3-319-50496-4_18.

[49] M. F. L. Schmitt and E. J. Spinosa, "Outlier detection on semantic space for sentiment analysis with convolutional neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Jul. 2018, pp. 1-8, doi:10.1109/IJCNN.2018.8489200.

[50] K.-Y. Chen, K.-W. Shih, S.-H. Liu, B. Chen, and H.-M. Wang, "Incorporating paragraph embeddings and density peaks clustering for spoken document summarization," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 207-214, doi:10.1109/ASRU.2015.7404796.

[51] Mary, "A density based dynamic data clustering algorithm based on incremental dataset," *Journal of Computer Science*, vol. 8, no. 5, pp. 656-664, May 2012, doi:10.3844/jcssp.2012.656.664.

[52] D. Cheng, Q. Zhu, J. Huang, Q. Wu, and L. Yang, "A hierarchical clustering algorithm based on noise removal," *Int. J. Mach. Learn. & Cyber.*, vol. 10, no. 7, pp. 1591-1602, Jul. 2019, doi:10.1007/s13042-018-0836-3.

[53] B. S. Harish, M. B. Revanasiddappa, and S. V. A. Kumar, "A modified support vector clustering method for document categorization," in *2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA)*, Singapore, Singapore, Sep. 2016, pp. 1-5, doi:10.1109/ICKEA.2016.7802982.

[54] J. Hou and M. Pelillo, "A new density kernel in density peak based clustering," in 2*016 23rd International Conference on Pattern Recognition (ICPR),* Cancun, Dec. 2016, pp. 468-473, doi:10.1109/ICPR.2016.7899678.

[55] M. Patidar, S. Rohatgi, A. Chaudhary, M. P. Singh, P. Agarwal, and G. Shroff, "Activity detection from email meta-data clustering," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW),* Barcelona, Spain, Dec. 2016, pp. 568-575, doi:10.1109/ICDMW.2016.0087.

[56] J. Lu and Q. Zhu, "An effective algorithm based on density clustering framework," *IEEE Access*, vol. 5, pp. 4991-5000, 2017, doi:10.1109/ACCESS.2017.2688477.

[57] D. Yu, G. Liu, M. Guo, X. Liu, and S. Yao, "Density peaks clustering based on weighted local density sequence and nearest neighbor assignment," *IEEE Access*, vol. 7, pp. 34301-34317, 2019, doi:10.1109/ACCESS.2019.2904254.

[58] M. Du, S. Ding, and H. Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," *Knowledge-Based Systems*, vol. 99, pp. 135-145, May 2016, doi:10.1016/j.knosys.2016.02.001.

[59] D. Wu, Y. Zeng, and Y. Qu, "Text document clustering based on density k-means," *dtcse, no. cmee*, Jan. 2017, doi:10.12783/dtcse/cmee2016/5349.

[60] P. H. Ahmad, and Shilpa Dang, "Performance evaluation of clustering algorithm using different datasets," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 3, no. 1, January 2015 pp. 167-173, 2015.