RESEARCH ARTICLE

Enhancing web search result clustering model based on multiview multirepresentation consensus cluster ensemble (mmcc) approach

Ali Sabah, Sabrina Tiun*, Nor Samsiah Sani, Masri Ayob, Adil Yaseen Taha

Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

* sabrinatiun@ukm.edu.my

Citation: Sabah A, Tiun S, Sani NS, Ayob M, Taha AY (2021) Enhancing web search result clustering model based on multiview multirepresentation consensus cluster ensemble (mmcc) approach. PLoS ONE 16(1): e0245264. https://doi.org/10.1371/journal.pone.0245264

Editor: Gulistan Raja, University of Engineering & Technology, Taxila, PAKISTAN

Received: September 23, 2020

Accepted: December 26, 2020

Published: January 15, 2021

Data Availability Statement: 1. MORESQUE dataset available at http://lcl.uniroma1.it/moresque/) 2. ODP-239 Dataset available at http://search.fub.it/odp-239/.

Funding: This research was funded by the Universiti Kebangsaan Malaysia (Grant code: GUP-2020-063). This grant was received by Dr Sabrina Tiun. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Existing text clustering methods utilize only one representation at a time (single view), whereas multiple views can represent documents. The multiview multirepresentation method enhances clustering quality. Moreover, existing clustering methods that utilize more than one representation at a time (multiview) use representation with the same nature. Hence, using multiple views that represent data in a different representation with clustering methods is reasonable to create a diverse set of candidate clustering solutions. On this basis, an effective dynamic clustering method must consider combining multiple views of data including semantic view, lexical view (word weighting), and topic view as well as the number of clusters. The main goal of this study is to develop a new method that can improve the performance of web search result clustering (WSRC). An enhanced multiview multirepresentation consensus clustering ensemble (MMCC) method is proposed to create a set of diverse candidate solutions and select a high-quality overlapping cluster. The overlapping clusters are obtained from the candidate solutions created by different clustering methods. The framework to develop the proposed MMCC includes numerous stages: (1) acquiring the standard datasets (MORESQUE and Open Directory Project-239), which are used to validate search result clustering algorithms, (2) preprocessing the dataset, (3) applying multiview multirepresentation clustering models, (4) using the radius-based cluster number estimation algorithm, and (5) employing the consensus clustering ensemble method. Results show an improvement in clustering methods when multiview multirepresentation is used. More importantly, the proposed MMCC model improves the overall performance of WSRC compared with all single-view clustering models.

1. Introduction

Information retrieval has become difficult due to the abundant online information currently available. Modern search engines respond to user queries with many results, and only a few prove to be relevant. Therefore, users are frequently compelled to browse through long lists of

Competing interests: The authors have declared that no competing interests exist.

results. Web search engines perform excellently, but a typical user often ends up browsing numerous pages to find his needs. Research continues with numerous advances to make the process relatively fast and easy. Clustering strategies could be used by search engines to group the results of searches into various categories. In this manner, users may navigate directly to a category and find what they need faster than by browsing pages with a traditional web search engine. The abundance of the huge, continually increasing amount of information makes information retrieval an essential, difficult process. As the amount of available information increases daily, the amount of search results returned by a user query also increases. Nowadays, search result responds to user queries, of which only a few are relevant. Clustering is an effective technique to organize retrieval results in a meaningful cluster. However, existing clustering techniques have several drawbacks. Web search result clustering (WSRC) brings several challenges. Thus, various studies have proposed adopting clustering methods [1], and several clustering algorithms have been proposed. Based on the literature [1, 2] and the results obtained by [3], no single clustering algorithm is appropriate for all types of data. No clustering method is suitable for all tasks, probably due to the different dynamics and nature of web search result datasets. The different clustering results are likely due to non-normally distributed data [1, 4, 5]. Combined different clustering algorithms in which the final clustering is selected from the candidate clustering solution with the maximum similarity from all candidate clustering solutions. Therefore, clustering algorithm combinations, such as the single clustering algorithm, are inappropriate for all data types, and existing clustering methods utilize only one representation at a time (single view). By contrast, documents can be represented by multiple views.

In addition, existing clustering methods that utilize more than one representation at a time (multiview) use representations with the same nature. Hence, using multiple views that represent data in totally different representations with clustering methods and create a diverse set of candidate clustering solutions is reasonable. Moreover, the poor performance obtained by the existing research may be because various user queries can result in a different number of clusters in searching clustering results [1, 4].

On this basis, an effective, dynamic short-text clustering method must consider combining multiple views of data including semantic view, lexical view (word weighting), and topic view as well as the number of clusters. The key contributions of this study are as follows: First, an enhanced multiview multirepresentation clustering model that combines different views and representation methods is proposed. The different view and representation methods include n-gram weighting, word embedding Word2vec, and Dirichlet multinomial mixture (DMM) topic to generate different candidate clustering solutions. Second, an efficient radius-based cluster number estimation algorithm is proposed for the dynamic estimation of the number of clusters in WSRC and to overcome the limitations of k-based clustering, in which the number of clusters must be predefined. Third, a multiview multirepresentation consensus clustering ensemble method (MMCC), is proposed.

The proposed method can be adapted for short text clustering from online social media platforms, such as Twitter, to help extract knowledge. Moreover, the idea of the proposed method in this research can be applied to data with different views. The proposed method can also be applied for data that refer to different representations to achieve a better performance compared with single-view data representation. Thus, it can be implemented in real-time WSRC applications.

This study is organized as follows: Section 2 provides an overview of the relevant studies. Section 3 presents the proposed method of multiview multirepresentation approach with consensus clustering (CC) ensemble to create different candidate solutions. Sections 4, 5, and 6 presents the results and discusses the conducted experiments. Section 7 concludes the study.

2. Related studies

The multiview clustering approach has drawn considerable attention in recent years, and it aims to exploit consensus and complementary information through multiple views [6]. Clustering of search results has become one of the most important strategies for extracting knowledge online [7]. Clustering of search results that typically return snippets of web results in thousands is often impractical. The key is to identify optimal universal stopping criteria for every query. Several researchers have suggested clustering methods that can resolve the problems of conventional clustering strategies and handling WSRC difficulties. In this section, an overview of advanced studies related to WSRC is provided. The issue has been addressed through diverse strategies, but the median partition-based strategy remains the most well-known approach for ensemble clustering thus far [1, 8–10]. A median partition-based strategy involves a lone candidate clustering solution that features maximum similarity among all clustering solution candidates. This strategy is selected as the ultimate clustering solution for this research.

Abdulameer et al. [3] introduced a new framework that enhances the performance of WSRC following the new wiki-based K-nearest neighbor (KNN) representation technique. The approach also offers a new unsupervised distributed word representation model, where every word is characterized by a vector of its semantically linked referents, including schematic expansions of user queries and word snippets. Using the new wiki- and KNN-based representation technique is recommended for resolving short-text problems and improving the general performance of WSRC. Lastly, the results of the suggested model are reviewed and compared with those of baseline methods to confirm the importance of the proposed WSRC approach.

Feng et al. [11] developed a new text clustering ensemble algorithm derived from semantic sequence algorithmic methods proposed for this research. Text clustering results from the application of K-means and semantic sequence algorithms were generated and combined in accordance with overlap coefficient similarity principles to produce the coassociation matrix among semantic sequences. Lastly, the final clusters were acquired by merging documents that correspond to similar semantic sequences in this matrix.

Wang [12] introduced an enhanced non-negative matrix factorization (NMF) algorithmic method to combine multiple clusters. First, the K-means algorithm was performed to partition the adjacent matrix of the hypergraph and acquire the indicator matrix, which was then imparted to NMF as the initial factor matrix. Second, NMF was conducted to obtain the basis and coefficient matrices. Lastly, clustering results were acquired using the elements in the coefficient matrix.

Abu-Jamous et al. [13] developed a new clustering paradigm based on a new binarization of the consensus partition matrix method. The technique exploits the results of multiple clustering tests across the same dataset to generate a single fuzzy consensus partition. They suggested tunable methods that binarize this partition mirror biological reality, such that certain genes may be assigned to numerous clusters, whereas others may not be assigned at all. This technique can express the relative tightness of multiple clusters to generate tight or wide overlapping clusters as well as extract the unique genes bearing the profiles of these multiple clusters concurrently.

Gravitational ensemble clustering (GEC) denotes an ensemble clustering technique developed by Sadeghian and Nezamabadi [8]. This approach uses ensemble and gravitational clustering principles to obtain superior clustering results. Thus, the GEC model uses several runs of the K-means algorithm using different parametric initializations to generate the set.

In recent years, the multiview clustering approach brought about a vital expansion of ensemble clustering methods. Clustering algorithms can be applied to varying views of data to obtain dissimilar cluster labels for similar sets of objects using the multiview clustering method. These methods are then combined such that the final clustering results generated are superior to the individual clustering results for each data multiview approach. Multiview clustering can be used in various steps of the proposed clustering model. Hussain et al. [14] introduced the multiview clustering algorithmic method, where different ensemble methods are combined for a better effect. The approach computes various similarity matrices for individual datasets and then aggregates these matrices to construct a combined similarity matrix that is subsequently used to acquire the final clustering results.

Wahid et al. [1] developed a new approach for cluster ensemble, namely, the multiview multiobjective cluster ensemble derived from the multiview multiobjective evolutionary algorithm. This approach accords with the development of crossover methods that produce new clusters for the candidate clustering solutions, the development of mutational techniques for margin clusters and splitting, and the improvement of multiobjective fitness functions for resolving multiobjective optimization problems. The main limitation of this strategy is that its implementation requires predefined multiple views. It also requires the investigation of automatic detections of multiview results in documents.

Boongoen & Iam-On [15] reviewed cluster ensemble approaches while covering new extensions and applications for a range of ensemble-generating strategies with summarizations and representations of ensemble members and discussion of the issues of CC. Their survey also included differing extensions and applications of cluster ensemble methods that focus on numerous research problems and challenges. The outcome obtained from the literature review of web search results showed that clustering is the most effective approach and an issue that encourages new solutions to increase effectiveness and efficiency with massive, heterogeneous, and dynamic web pages. However, despite much research in the field of web search result, several open issues, such as the achievement of a better quality and an effort to deal with the drawbacks of each technique, need further investigation. At present, the majority of multiview clustering methods are derived from single-view concepts, where one measure of quality for partitioning is optimized explicitly or implicitly using various paradigms for single-view learning. Current clustering techniques use only single representations at any time (single views), whereas documents can be represented by multiple views. Moreover, multiple representations may bring about multiple cluster sets and provide multiple insights into the data.

Multiview clustering has been successfully applied to numerous applications including computer vision [16–18], natural language processing [19, 20], social media [21–23], bioinformatics, and health informatics [24–26]. Thus, an efficient, dynamic short-text clustering technique should consider the combination of multiple views of data because different views in parallel have been proven to improve clusters [27]. Thus, using multiple views for producing multiple candidates in a clustering collection technique is logical. Therefore, in this study, a method to enhance the clustering techniques with a new multiview, multirepresentation clustering framework that integrates n-gram weighting, DMM, and word embedding Word2vec to generate multiple candidate clustering solutions is proposed. The following section describes the framework for enhancing the WSRC model based on our proposed method, the MMCC.

3. Proposed multiview multirepresentation cluster ensemble model

The section presents the detailed description of the framework for enhancing clustering models for WSRC based on the proposed multiview multirepresentation and the ensemble consensus clustering method (MMCC). The framework consists of six stages. Fig 1 describes the framework to develop the MMCC ensemble method and enhance the WSRC model.

The proposed framework includes several stages. The first stage is identifying and selecting appropriate datasets to be used. The datasets, Open Directory Project (ODP)-239 and

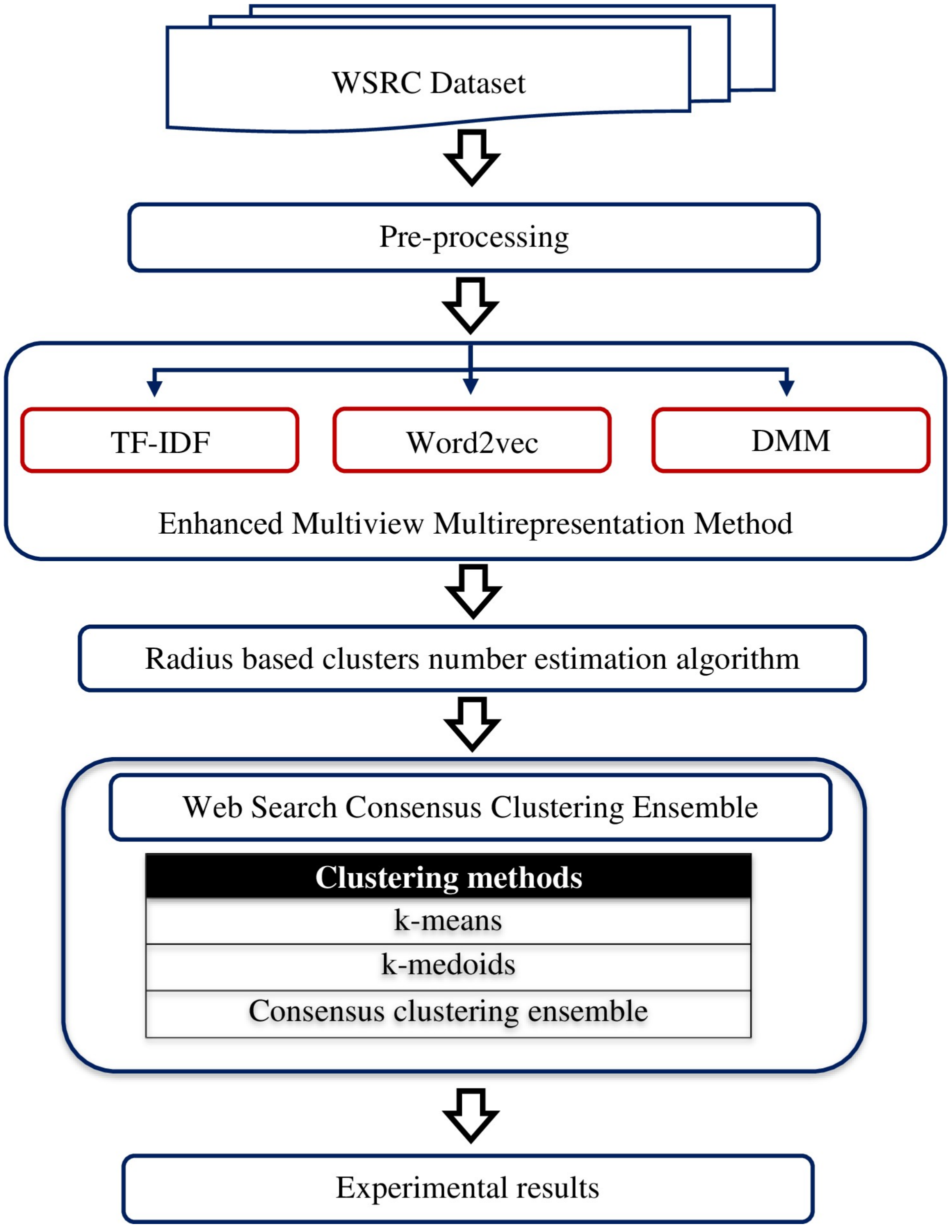
WSRC Dataset
Pre-processing
TF-IDF
Word2vec
DMM
Enhanced Multiview Multirepresentation Method
Radius based clusters number estimation algorithm
Web Search Consensus Clustering Ensemble
Clustering methods
k-means
k-medoids
Consensus clustering ensemble
Experimental results

Fig 1. Proposed framework for enhanced WSRC model.

https://doi.org/10.1371/journal.pone.0245264.g001

MORESQUE, which are selected in our study to validate the clustering methods, were extensively used in several previous studies [28–32]. The second stage involves applying the preprocessing techniques for the datasets. The third stage is applying our proposed multiview multirepresentation clustering method, a combination of different views and representation methods (n-gram weighting, word embedding Word2vec, and DMM topic), to create different candidate clustering solutions. The fourth stage uses an efficient radius-based cluster number estimation algorithm to determine the dynamic number of clusters. The fifth stage proposes the MMCC method by combining different views and representation methods to generate various candidate clustering solutions. The last stage presents the standard evaluation metric to evaluate the proposed method.

3.1 Multiview multirepresentation for clustering models

Diverse representations result in different sets of clusters and may provide different data insights, considering that the multiview multirepresentation method was already proven to enhance clustering quality [27]. Recently, the trend moved toward formulating clustering ensembles as optimization problems to enhance the results [33, 34]. Hence, using multiple views that represent data in totally different representations with clustering methods and create a diverse set of candidate clustering solutions is reasonable.

Multiview data refer to different representations of the same data instance. Each view may have diverse statistical properties that result in various insights of the data. Compared with previous clustering ensemble methods [1, 2, 35], this study adopts views and representations that are completely different to generate initial candidate clustering solutions. Thus, a multiview multirepresentation of data is introduced to create a clustering ensemble suitable for large data types. These representations and views ensure that different candidate clustering solutions are created by various clustering methods. In addition, cluster ensemble results are remarkably affected by the level of document representation [36]. Moreover, different representations may capture varying degrees of explanatory ingredients hidden in short documents. This study aims to enhance the performance of the clustering solution by selecting high-quality clusters from diverse candidate clustering solutions created by different clustering methods. From this perspective, several views of data are considered in the proposed approach, including n-grams, senses, topics of snippets, user query, and titles. In addition, three diverse representations, namely, term frequency–inverse document frequency (TF–IDF) matrix representation, word embedding representations (Word2vec), and latent topic representation (DMM), are used. The proposed multiview multirepresentation is shown in Fig 2.

These representation methods are selected due to the following reasons:

1. All representations have different statistical properties or formulations and come from diverse genres to ensure varied candidate clustering solutions.

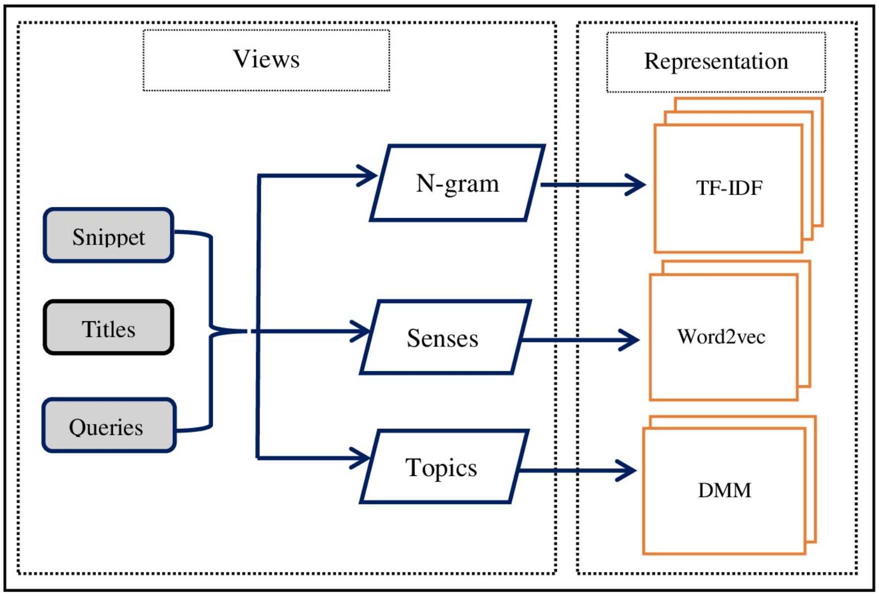

Fig 2. Proposed multiview multirepresentation models.

https://doi.org/10.1371/journal.pone.0245264.g002

2. Methods such as DMM derived from Latent Dirichlet Allocation are popular and still among the most effective representation methods [37, 38].
3. The most evident advantage of DMM and Word2vec representations is that they reduce the dimensions of features for representing a document [2, 37, 39].
4. Word2vec is a modern representation method that efficiently captures the implicit semantics of words [2, 39]. DMM also captures the implicit semantics of words [40].

In n-gram with TF–IDF representation matrix, every row represents a search result document, every column signifies an n-gram, and a cell contains the TF–IDF weighted value of an n-gram for a search result document. In word embedding representations, each search result document signifies a set of n vectors. Every vector represents a word from the search result document, and each vector of a word contains its most *k* similar words with their similarity values obtained using Word2vec. In a DMM, rows represent documents, and N columns denote N topics. These representations are obtained from different genres, and the n-gram with TF–IDF matrix representation is based on a simple numerical statistic. The word embedding representation is a word representation based on co-occurrence statistics and deep neural network. DMM is based on multinomial probability distribution theory.

3.1.1 Query–snippet–title n-gram view and representation. In text clustering, a document is represented by a vector of terms and their weight values called attribute values. A common weighting scheme is assigned to each word based on TF–IDF [41–43]. A search engine generates a set of search results when using a query. Each search result consists of query, snippet, and title (QST). In query–snippet–title n-gram view and representation (QSTNVR), QST is combined and represents a bag of the n-gram. QST is tokenized into an n-gram list, which contains a list of unigram, bigram, and trigram. Each n-gram is weighted using the TF–IDF weighting scheme. A QST in n-gram is represented by Eq (1):

$$SR = \begin{bmatrix} TF - IDF(g_1, SR) \\ TF - IDF(g_2, SR) \\ TF - IDF(g_3, SR) \\ \vdots \\ TF - IDF(g_n, SR) \end{bmatrix}, \tag{1}$$

where n in g_n is the whole number of n-gram, and *TF–IDF* is the *TF–IDF* function defined by Eq (2):

$$TF - IDF(g_i, SR) = tf(g_i, SR) * \log\left(\frac{N}{df(g_i)}\right), \tag{2}$$

where TF is the term frequency in document d, *IDF* is the number of appearances of this term in the document, $tf(g_i, SR)$ is the frequency of n-gram g_i in the search result, |N| is the whole number of search, and $df(g_i)$ is the number of search results with n-gram g_i. QSTNVR representation uses the lexical representation of the search result document and neglects its semantic representations.

3.1.2 QST Word2vec view and representation. Representations of word embeddings, such as Word2vec, are effective in describing fine-grained semantic associations between words [44]. Distributed word representation, Word2vec, interprets low-dimensional, closely packed word embeddings and efficiently captures the implicit semantics of words [2, 39, 40, 45, 46]. The low-dimensional embedding of words is much more appropriate for the latest deep learning neural-based models compared with conventional one-hot representation. Based on word embeddings, Word2vec represents every word denoted by $\boldsymbol{x}$ in a vocabulary, denoted by $\mathbf{V}$ as a low-dimensional compact vector, and denoted by $\mathbf{v_x}$ in space $\mathbb{R}^{\mathbf{D}}$. Such word representations can be learned based on the distributional concept, which assumes that words with related contexts are likely to have similar sense. Skip-gram's training goal is to search for word representations, which are helpful in predicting the neighboring words in a statement from a large textual collection. Given a series of training words represented by $\{\boldsymbol{\omega_1}, \boldsymbol{\omega_2}, \ldots, \boldsymbol{\omega_T}\}$, the goal of the skip-gram model is to increase the mean log probability in Eq (3) [44, 47]:

$$J(\Omega) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(\omega_{t+j}|\omega_t), \tag{3}$$

where $\boldsymbol{\Omega}$ represents the parameters of the model to be trained (defined later), c is the training context size (a sliding window surrounding the center word ω_t), and $p(\omega_{t+j}|\omega_t)$ denotes the probability of finding the word ω_{t+j} with respect to the center word ω_t and is defined by a basic hidden layer neural network model. The network has three strata, which include a hidden layer, an input layer, and a SoftMax output layer, corresponding to the words in the context window. Generally, the network serves as the input $\omega_t \in \mathbb{R}^V$, where V represents the size of the vocabulary. It creates a hidden state represented by $h \in \mathbb{R}^D$, where D denotes the hidden layer size or embedding space dimension, which is further converted to the output represented by $\omega_{t+j} \in \mathbb{R}^V$. Multiple layers are completely connected, where the weight matrix M_{out} at the output strata is shared among all the contextual words. Gathering these weight matrices from this model represents the parameter of the model by $\boldsymbol{\Omega} = \{M_{in}^{V\times D}, M_{out}^{V\times D}\}$. In the skip-gram architecture based on Eq (3), the input is represented by ω_t functions as a sparse vector of a one-hot (or 1-of-V) encoding where the element related to input ω_t is 1, and the remaining components are set to 0. Thus, the basic formulation of skip-gram defines $p(\omega_{t+j}|\omega_t)$ as the function of SoftMax in Eq (4) [47]:

$$p(\omega_O|\omega_I) = \frac{\exp(\langle V_{in}^{\omega_I}, V_{out}^{\omega_O}\rangle)}{\sum_{\omega=1}^{V}\exp(\langle V_{in}^{\omega_I}, V_{out}^{\omega_O}\rangle)}, \tag{4}$$

where $\langle .,. \rangle$ represents the inner product amongst two vectors, V_{in}^{ω} and V_{out}^{ω}, which are the "input" and "output" vectors of ω, respectively, and relate to the rows of the model parameter matrices denoted by M_{in} and M_{out}.

3.1.3 QSTDMM view and representation. The DMM model used in this study [37, 38, 48], a popular topic modeling method that is used to extract a topic from the documents, has semantic information. DMM typically refers to the mixture of the unigram model. In applying the DMM model for search result clustering purposes, QST is assumed to belong to one topic. Each snippet is associated with at most one topic. Corpus is a set of search results composed of $\boldsymbol{D}$ QSTs. $\boldsymbol{D}$ is the number of QST in the corpus. Each QST → $\mathbf{d}$ contains a set of words ($\mathbf{w}\epsilon\mathbf{1,2,..,N_d}$). The DMM model D in Fig 3 has the following generative process:

a. $\alpha = (\propto_1, \propto_2, \ldots, \propto_n)$ is sampled from a Dirichlet distribution with parameter $\lambda = (\lambda_{\propto 1}, \lambda_{\propto 2}, \ldots, \lambda_{\propto n})$.

b. For each topic $t = 1,2,\ldots,n_T$, β_t is sampled from a Dirichlet (λ_β), where $\beta = (\lambda_{\beta 1}, \lambda_{\beta 2}, \ldots, \lambda_{\beta n})$.

c. For a QST → d ($d\epsilon\{1,\ldots,$N$\}$): in QST d,

   1. A topic $z_n \in \{1,2,\ldots,T\}$ is selected from Multinomial(α) where $\alpha = (\propto_1, \propto_2, \ldots, \propto_n)$ represents the topic distribution in the corpus.

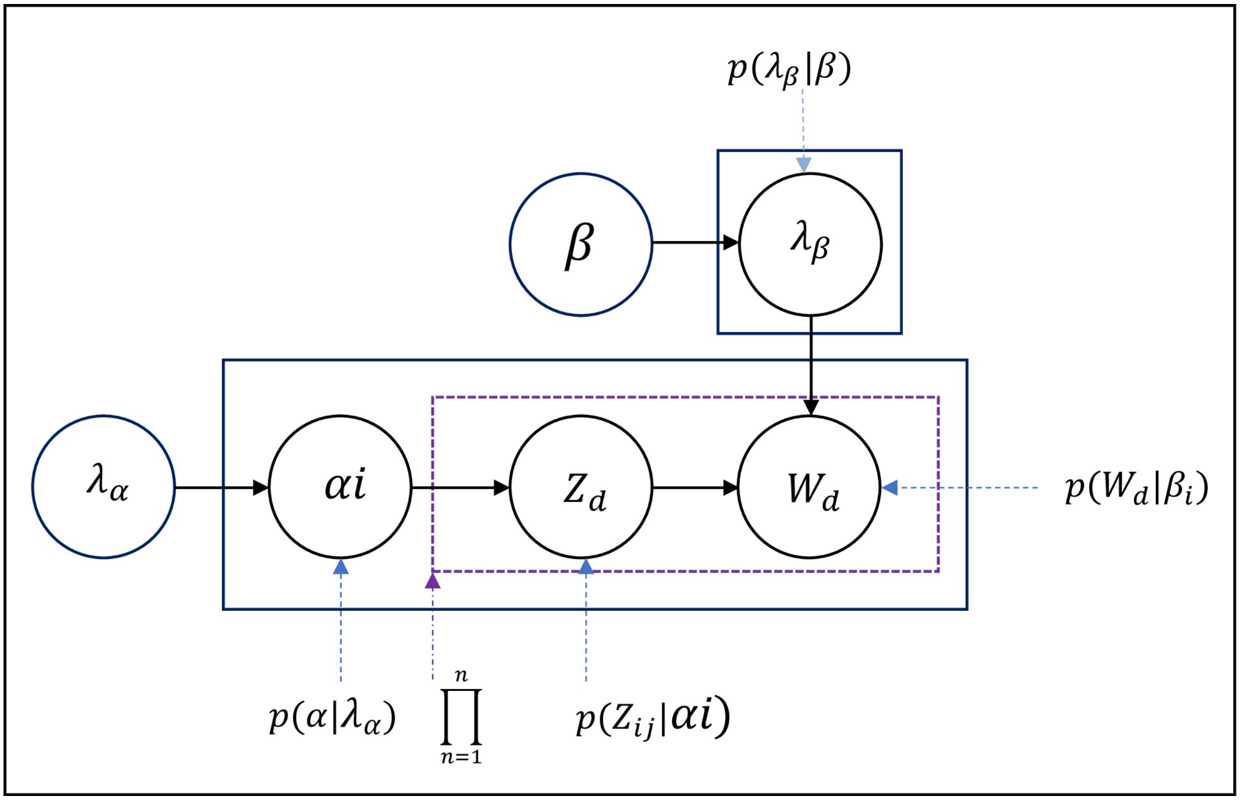

Fig 3. DMM models in the generative process.

https://doi.org/10.1371/journal.pone.0245264.g003

2. Word count N_d is selected, and a word w_d from QST d from Multinomial(β) is independently selected, where $\beta = (\beta_{\propto 1}, \beta_{\propto 2},. . .,\beta_{\propto n})$ represents the word topic distribution in the corpus.

The probability of observed data D is computed and maximized in Eq (5) to infer the latent variables and hyperparameters [16, 37]:

$$p(d|\alpha, \beta) = \sum_{l=1}^{T} \alpha_l \frac{N_d!}{\prod_{w=1}^{V} N_d^w!} \prod_{w=1}^{V} \beta_{wt}^{N_d^w} , \tag{5}$$

where α indicates the parameters of topic Dirichlet prior and the distribution of words over topics from the Dirichlet distribution given β, V is the size of the vocabulary, N_d is the number of words in the QST, and N_d^w is the frequency of word w in QST d. The Dirichlet-multinomial pair for the corpus-level topic distributions is (α, θ).

In a QSTDMM output matrix, rows represent documents, and columns denote topics. A cell (i,j) is labelled 1 if document d_i is from topic t_j. The most evident advantage of this representation is reducing the dimensions of features for representing a document.

3.2 Radius-based number of cluster estimation algorithm

WSRC is a dynamic clustering. Each query has its own search result documents. K-based clustering algorithms are more challenging in search result clustering, where every user's queries can result in a different number of clusters. This study introduces a radius-based number of the cluster's estimation algorithm to overcome the limitations of k-based clustering, where the number of clusters must be predefined. The stepwise flow of the proposed algorithm is summarized as follows (see Algorithm 1):

Step i: Inferring a search result graph: Given a query and its search result documents $Q_d = \{r_1, \ldots, r_m\}$, each document is represented as a node. The edge between two nodes e_{ij} is straightforwardly computed between each pair of search result nodes (r_i, r_j) given a radius value β, which is a random number less than ≤ 0.5 using Eq (6) [49]:

$$we_{ij} = \begin{cases} \cos(r_i, r_j) = \dfrac{|r_i| * |r_j|}{\sqrt{r_i^2 * r_j^2}} & if\ \cos(r_i, r_j) > \beta \\ 0 & otherwise \end{cases} \tag{6}$$

Step ii: Density and degree estimation: The density of each search result document r_i from the dataset is defined as the sum of the weight of all its edges and calculated using Eq (7) [49]:

$$Density(ri) = \sum\nolimits_{\mathrm{k}} (e_{ik}) \tag{7}$$

Step iii: Highest density document selection: The document or node ri, which has the highest density (Density(ri)>Density(rj); ∀j) is selected as the first seed S_{j1}. All nodes connected to it are removed from the dataset and added to its cluster.

Step iv: Q optimal target cluster selection: Steps (iii) and (iv) are repeated to continue the selection of the subsequent highest density *document* as long as the graph still has nodes.

```
Algorithm 1: Radius-based number of cluster estimation algorithm
Input: Given a query and its search result documents $Q_d = \{r_1, \ldots, r_m\}$
Output: Set of candidate centroids C
BEGIN
  Step i: The graph similarity matrix WE is calculated
  For each document pairs ($r_i$, $r_j$) in $Q_d$
    $\beta$←select_small_random_number(0.50)
   Begin
      sim←calclate_cosine_similarity($r_i$, $r_j$)
        if(sim>$\beta$)
    WE[i,j]←sim
          Else if
    WE[i,j]←0
      End if
   End
  Step ii: The document with the highest density is identified
  Density_max←0
    For i = 1 to n // n number of search results documents
   Begin
      Doc_density←0
        For j = 1 to n // n number of search results documents
          Begin for
         $Doc_{density}$←$Doc_{density}$+WE[i,j]
         End for
  Step iii: The document with the highest density is selected
   if ($Doc_{density}$>Density_max)
      Begin if
   Density_max←$Doc_{density}$
        $Dens_{document}$←ri// remove all documents connected with
$Dens_{document}$
                  by making all of its row 0
```

```
            For k = 1 to n // n number of search results documents
                        If (WE[i,k]>0)// if document rk is connected with document ri
                          Begin if
                              For v = 1 to n // n number of search result documents
                                    WE[i,k]←0
                            End if
                    End if
            End
    **Step iv**: **The document with the highest density is added to candidate centroids, and this process is repeated**
            **Begin**
        C←{Dens_document}
        repeat step 2 and 3 until all entries in matrix WE are zeros
          End
     End
return C
```

3.3 Consensus clustering ensemble method

Cluster ensemble methods are learning techniques that create a cluster set to produce new information points by considering their predictions' vote. A sufficient, necessary condition to obtain a more precise cluster collection than any individual member is that the clusters must be separate and accurate [50]. The chief finding is that such collections are often very precise than their individual clusters, and the combination of many single strategies improves the performance of the ultimate classifier reliability, accuracy, and understandability [51]. In addition to regression and categorization, ensemble techniques have been developed for clustering as well as other machine learning functions [9].

Collecting various clustering solutions is recommended, where diversity is produced by allowing every base clustering to be created using different clustering techniques. The main idea is that every clustering method has advantages and drawbacks, and is appropriate for certain types of datasets. Different clustering models can provide various decisions on clustering of data and support one another. Therefore, combining the outcomes of different clustering techniques can ensure better clustering of data. Multiple techniques can provide multiple decisions on data components and complement one another [1, 5, 34]. This study uses three views and representations, namely, QSTNVR, QSTWVR, and query–snippet–title DMM view and representation (QSTDVR) with k-based clustering methods, namely, k-means and k-medoids, to generate the ensemble. In addition, base clustering must be precise and varied to acquire a high-quality ensemble [9, 52, 53].

Cluster ensembles are effective for improving the robustness and stability of unsupervised clustering solutions [1, 34, 35, 54]. However, finding a CC to combine multiple-base clustering is the main objective of this study. In Algorithm 2, CC is applied on all solutions to generate a set of clustering solutions $p = \{p_1, p_2,\ldots.., p_m\}$, where each clustering solution $p_\mathrm{i} = \{C_1,\ldots.., C_K\}$. These different clustering solutions are created by k-means and k-medoids clustering algorithms with multiview and multirepresentation. QST, which agrees with m clustering solution, should be determined to achieve a consensus among P clustering methods (with clustering solutions). First, let u and v be two QSTs of X, *where* $X = \{x_1, x_2,\cdot\cdot\cdot, x_n\}$ is a set of n QST. Then, the algorithm determines the frequency when QST u and v are clustered together in all clustering methods P. The algorithm determines the frequency when two objects are clustered given two clustering solutions p1 and p2 produced by two different clustering methods. Let T be the similarity measure, as follows [55]:

$$T_{u,v}(p_i,p_j) = \begin{cases} 1 & \textit{if } p_i(u) = p_i(v) \textit{ and } p_j(u) = p_j(v) \\ 0 & \textit{otherwise} \end{cases}$$

$S_{u,v}(p_i, p_j) = \Sigma_{i,j} T_{u,v}(p_i, p_j)$, where u and v are consecutive QSTs in x. $S_{u,v}$ is the sum of all similarities between clustering method (clustering solution) p_i and clustering method p_j. A consensus is reached between QST u and QST v when their corresponding vectors satisfy $S_{u,v}(p_i, p_j)/m>0.5$, that is, when a majority of clustering models agree, as shown in Step 1. This finding indicates that more than half of the clustering methods assign u and v QSTs to the same cluster, as shown in Step 2. The CC ensemble algorithm is illustrated in Algorithm 2.

```
Algorithm 2: Consensus Clustering Ensemble Method
Input: Set of clustering solutions $p = \{p_1, p_2, ...., p_m\}$
Output: Final clustering $P$
BEGIN
Step i: find agreement $\alpha$ btween clustering solutions $p = \{p_1, p_2, ...., p_m\}$
for each QST pair
    For all QST pairs ($d_i$, $d_j$)
    Begin
    $v_i \leftarrow \{p_1(d_i), p_2(d_i), ...., p_m(d_i)\}$
    $v_j \leftarrow \{p_1(d_j), p_2(d_j), ...., p_m(d_j)\}$
            For each clustering solution $p_z$ do
          If ($p_z(d_i)$ = = $p_z(d_j)$)
        $\alpha = \alpha+1$;
    End if
End for
If ($\frac{\alpha}{m} > 0.5$) // m is the number of clustering solutions
    Step ii: if more than half of the methods uses $d_i$, $d_j$ in the same
cluster, then create cluster $C_u$ and place them in it
    Create $C_u$ and $C_u \leftarrow \{d_i, d_j\}$
    Else
    Step iii: else if more than half of the methods uses $d_i$, dj in dif-
ferent clusters, then create clusters $C_u$ and $C_a$ and place each in dif-
ferent clusters
    $C_a \leftarrow d_i$
    Create $C_b$ and $C_b \leftarrow \{d_j\}$
    End if
    End for
End
```

4. Experimental results

4.1 Dataset

Clustering search engines use WSRC methods that integrate documents and provide descriptions for the collected documents. Thus, users find the required results effectively. The experiments are devised to use freely available text datasets and compare baseline models with recent clustering techniques. The datasets utilized for validation have been extensively used in several previous studies to validate the clustering techniques [29–31]. A range of gold standards has been used for investigating the search result clustering techniques, among which MORESQUE and ODP-239 are very popular. In the ODP-239 [28] technique, the documents are represented by a heading and a web summary, and the subtopics are selected from the uppermost layer of DMOZ6. The subtopics follow a normal distribution in MORESQUE because these subtopics are described depending on the uncertainty of Wikipedia [32]. Therefore, most

Table 1. Gold standard datasets of WSRC.

Dataset	No. of Queries	No. of Subtopics Average/Minimum/Maximum	No. of Snippets
ODP-239	239	10/10/10	25580
MORESQUE	114	6.7/2/38	11402

https://doi.org/10.1371/journal.pone.0245264.t001

query-related responses are included in the subtopics, but not all queries are related to Wikipedia or are vague. For instance, although the results of the search query produce several "world cup sports," the query containing the keywords "world cup" entered in Wikipedia cannot be regarded as vague. Table 1 presents an overview of two datasets.

4.2 Evaluation metrics

This study uses the standard information retrieval metrics precision, recall, and f-measure [56–59] to evaluate the proposed model. Precision intends to evaluate the cluster based on the number of correctly retrieved instances out of the total number of retrieved queries. Recall aims to evaluate the cluster based on the number of correctly retrieved instances out of the total number of correct instances in the dataset. Each cluster is treated as if it is the result of a query, and each class is the desired set of documents for a query. The precision, recall, and F-measure of the cluster are calculated as follows:

$$Precision_{ci} = \frac{\#\ of\ correct\ instance\ of\ cluster\ ci\ given\ a\ query}{total\ \#\ of\ cluster\ ci\ instances} \tag{8}$$

$$recall_{ci} = \frac{\#\ of\ correct\ instance\ of\ cluster\ ci\ given\ a\ query}{total\ \#\ of\ instances\ given\ a\ query} \tag{9}$$

$$Precision = \frac{\sum_{i=1}^{number\ of\ clusters} Precision(C_i)}{number\ of\ clusters} \tag{10}$$

$$recall = \frac{\sum_{i=1}^{number\ of\ clusters} recall(C_i)}{number\ of\ clusters} \tag{11}$$

Now, the f-measure can be calculated as follows:

$$f - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

5. Results

This section presents the results of the multiview multirepresentation on two k-based clustering methods, namely, k-means and k-medoids for WSRC. Other important results are from the conducted experiments of the proposed MMCC compared with single-view clustering methods. The following subsections present and explain the results in detail.

5.1 Results of multiview multirepresentation with individual clustering methods on ODP-239 dataset

This subsection presents the results of three views and representations, namely, QSTNVR, QSTWVR, and QSTDVR with k-based clustering methods, namely, k-means and k-medoids, for WSRC with the ODP-239 dataset.

Tables 2–4 show the results in terms of precision, recall, and f-measure value using the ODP-239 dataset. Each clustering method behaves differently under QSTNVR, QSTWVR, and QSTDVR. The performance of k-means differs from one view to another. Similarly, the performance of k-medoids varies from one view to another, that is, the f-measure values that represent the quality of clustering differ for the three views.

Tables 2–4 also show that QSTWVR outperforms QSTNVR and QSTDVR for both clustering methods (k-means and k-medoids). This finding may be due to the ability of QSTWVR to learn dense and low-dimensional word embeddings, and efficiently capture the implicit semantics of words; thus, it is more appropriate for short text [44].

Most importantly, with QSTWVR and QSTDVR, the best results obtained are those of k-means (with f-measures of 84.36 and 84.12, respectively), and the best result for QSTNVR is that of k-medoids with an f-measure of 75.3.

5.2 Results of the three multiview multirepresentation with clustering methods on the MORESQUE dataset

This subsection shows the results of QSTNVR, QSTWVR, and QSTDVR with k-means and k-medoids, for WSRC, on the MORESQUE dataset. Tables 5–7 present the results in terms of precision, recall, and f-measure on the MORESQUE dataset.

Table 2. Performance precision of QSTNVR, QSTWVR, and QSTDVR with k-means and k-medoids on the ODP-239 dataset.

Clustering Methods	QSTDVR	QSTWVR	QSTNVR
k-means	84.16	84.13	71.73
k-medoids	81.3	79.85	75.46

https://doi.org/10.1371/journal.pone.0245264.t002

Table 3. Performance recall precision of QSTNVR, QSTWVR, and QSTDVR with k-means and k-medoids on the ODP-239 dataset.

Clustering Methods	QSTDVR	QSTWVR	QSTNVR
k-means	83.12	84.20	71.07
k-medoids	78.03	82.45	71.05

https://doi.org/10.1371/journal.pone.0245264.t003

Table 4. Performance (f -measure) of QSTNVR, QSTWVR, and QSTDVR with k-means and k-medoids on the ODP-239 dataset.

Clustering Methods	QSTDVR	QSTWVR	QSTNVR
k-means	84.12	84.36	72.7
k-medoids	82.3	81.45	75.3

https://doi.org/10.1371/journal.pone.0245264.t004

Table 5. Performance (precision) of QSTNVR, QSTWVR, and QSTDVR with k-means and k-medoids on the MORESQUE dataset.

Clustering Methods	QSTDVR	QSTWVR	QSTNVR
k-means	84.66	84.76	75.46
k-medoids	81.16	83.65	77.07

https://doi.org/10.1371/journal.pone.0245264.t005

Table 6. Performance (recall) of QSTNVR, QSTWVR, and QSTDVR with k-means and k-medoids on the MORESQUE dataset.

Clustering Methods	QSTDVR	QSTWVR	QSTNVR
k-means	87.13	86.23	74.28
k-medoids	84.06	84.98	78.22

https://doi.org/10.1371/journal.pone.0245264.t006

Table 7. Performance (f-measure) of QSTNVR, QSTWVR, and QSTDVR with k-means and k-medoids on the MORESQUE dataset.

Clustering Methods	QSTDVR	QSTWVR	QSTNVR
k-means	86.01	85.73	75.26
k-medoids	83.06	84.69	77.32

https://doi.org/10.1371/journal.pone.0245264.t007

This section shows that the results of the multiview multirepresentation with individual clustering methods on ODP-239 dataset, k-means, and k-medoids behave differently with QSTNVR, QSTWVR, and QSTDVR. The performance of k-means and k-medoids differs from one view to another. Thus, the f-measure values that represent the quality of clustering are different for the three views.

However, Table 7 shows that QSTWVR and QSTDVR perform equally well (with f-measures of 83.06 and 84.69, respectively), and have better results than QSTNVR (f-measure = 77.32). Most importantly, for QSTWVR and QSTDVR, the best results are with k-means (f-measure values of 86.01 and 85.73, respectively). For QSTNVR, the best result is with k-medoids (f-measure = 75.26). However, QSTNVR has the worst performance using both clustering methods.

The comparative results of the clustering methods on both datasets show that clustering performance varies from dataset to dataset. However, the results obtained on the ODP-239 dataset are higher than those on the MORESQUE dataset due to the size, number of clusters per query, and different natures of the MORESQUE dataset. Moreover, different results can be obtained given the non-normally distributed data.

5.3 Results of MMCC ensemble

This section shows the results of MMCC ensemble on the ODP-239 and MORESQUE datasets. The CC ensemble method is constructed by combining three views and representations. V1, V2, and V3 (as shown in Tables 8 and 9) are used to denote QSTNVR, QSTWVR, and QSTDVR, respectively, for simplicity.

1. QSTNVR **(V1)**
2. QSTWVR **(V2)**
3. QSTDVR **(V3)**

Table 8. Performance of proposed MMCC on the ODP-239 dataset.

	V3C1		V3C2	
	V2C1	**V2C2**	**V2C1**	**V2C2**
V1C1	89.33	86.2	90.25	89.62
V1C2	89.92	89.62	91.15	87.17

https://doi.org/10.1371/journal.pone.0245264.t008

Table 9. Performance of proposed MMCC on the MORESQUE dataset.

	V3C1		V3C2	
	V2C1	**V2C2**	**V2C1**	**V2C2**
V1C1	88.04	85.26	89.68	88.67
V1C2	89.23	88.85	89.79	86.47

https://doi.org/10.1371/journal.pone.0245264.t009

For example, the column name "V3C1" in Tables 8 and 9 refers to the QSTDVR as V3 and k-means as C1.

All possible combinations use k-means (C1) and k-medoids (C2) methods. CC ensemble models must use all three views and representations to create a set of different candidate solutions and select high-quality clusters from the candidate solutions and form a superior clustering solution.

Table 8 displays the f-measure results of the proposed MMCC models on the ODP-239 dataset. All MMCC models outperform all previous clustering methods presented in Tables 4 and 6 for WSRC on the ODP-239 dataset. Table 9 displays the f-measure results of the proposed MMCC models on the MORESQUE dataset. The results show that MMCC models outperform the previous clustering methods in Tables 4 and 7 for WSRC on the MORESQUE dataset.

The results in Tables 8 and 9 also show that all MMCC models overcome the instability problem of single-view clustering models, and MMCC clustering models are stable and suitable for all types of data. In this study, "stable" is a reference on a method/algorithm that can achieve the best results on all given datasets.

Fig 4 shows that the results of MMCC models improve compared with those of all single-view clustering algorithms.

6. Discussion

This study investigates the performance of the overall multiview multirepresentation with conventional clustering technique, such as k-means and k-medoids, for WSRC and determines which algorithm provides better results. This study also presents the performance of proposed multiview multirepresentation clustering models that combine n-gram weighting, word embedding Word2vec, and DMM to generate different candidate clustering solutions. The incorporation of the proposed multiview multirepresentation based on consensus clustering ensembles (MMCC) enhances the performance of WSRC. The study aims to create a set of different candidate solutions and select high-quality clusters to form a superior clustering solution. Subsequently, the proposed MMCC improves the overall performance of WSRC compared with that of single-view clustering algorithms, as shown in Fig 4.

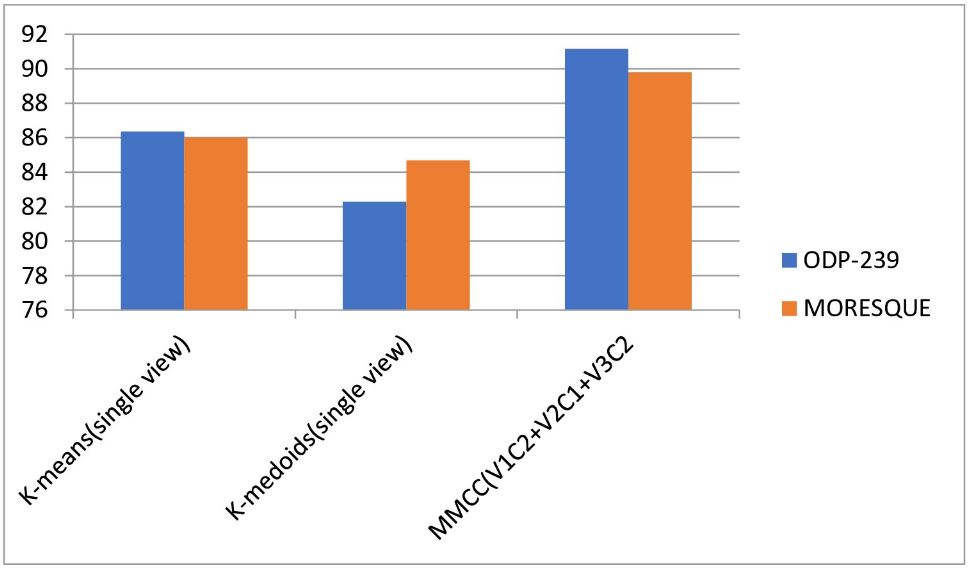

Fig 4. Performance comparison of the single view and MMCC.

https://doi.org/10.1371/journal.pone.0245264.g004

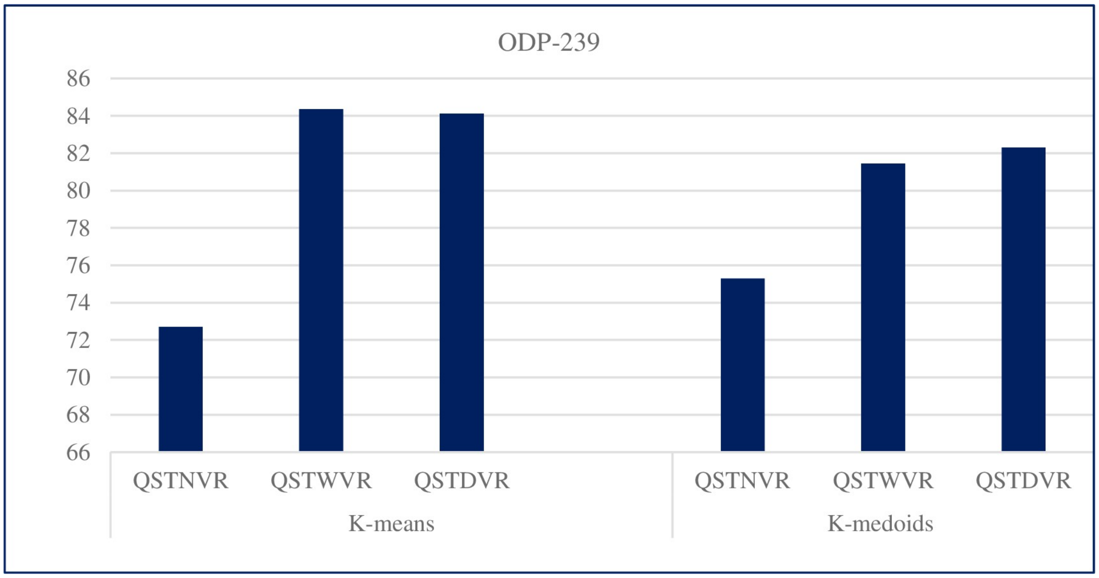

Fig 5. Performance of the three single-view models with k-means and k-medoids on the ODP-239 dataset.

https://doi.org/10.1371/journal.pone.0245264.g005

This study investigates the effect of the proposed MMCC for WSRC on two standard datasets. Investigation is performed by comparing single-view clustering models with the MMCC model. Single-view clustering models, which use k-means and k-medoids, achieve good clustering performances. The performance of single-view clustering results obtained using the ODP-239 dataset with QSTNVR, QSTWVR, and QST-DVR with k-means are 72.7, 84.36, and 84.12, respectively. Similarly, the performance of single-view clustering results obtained using the ODP-239 dataset with QSTNVR, QSTWVR, and QST-DVR with k-medoids are 75.3, 81.45, and 82.3 respectively. Fig 5 displays the overall performance comparison of the three single-view models with k-means and k-medoids on the ODP-239 dataset.

By contrast, using k-means for the ODP-239 dataset shows a high result of 84.36 in terms of f-measure with QSTWVR. This finding may be due to the ability of QSTWVR low-dimensional word embeddings to capture the implicit semantics of words efficiently; thus, it is more appropriate for short text [44, 60]. Likewise, the ODP-239 dataset result obtained with k-medoids is 82.3 in terms of f-measure using QSTDVR, which is also high compared with other clustering methods.

The performances with single-view clustering results on the MORESQUE dataset using k-means based on three views, QSTNVR, QSTWVR, and QSTDVR, are 75.26, 85.73, and 86.01, respectively. Similarly, the performances of k-medoids with three views of QSTNVR, QSTWVR, and QSTDVR are 75.3, 81.45, and 82.3, respectively. Fig 6 displays the overall performance of the three single-view models with k-means and k-medoids on the MORESQUE dataset.

By contrast, using k-means obtains a high f-measure of 86.01 on the MORESQUE dataset using QSTDVR. The f-measure obtained using k-medoids is 82.03 using QSTDVR, which also achieves high results compared with other clustering methods. The comparative results of the clustering methods on both selected datasets demonstrate that clustering performance varies from dataset to dataset. However, the results on the ODP-239 dataset are higher than those on the MORESQUE dataset due to different results that can be obtained given the non-normally distributed data, size, number of clusters per query, and different natures of the MORESQUE dataset.

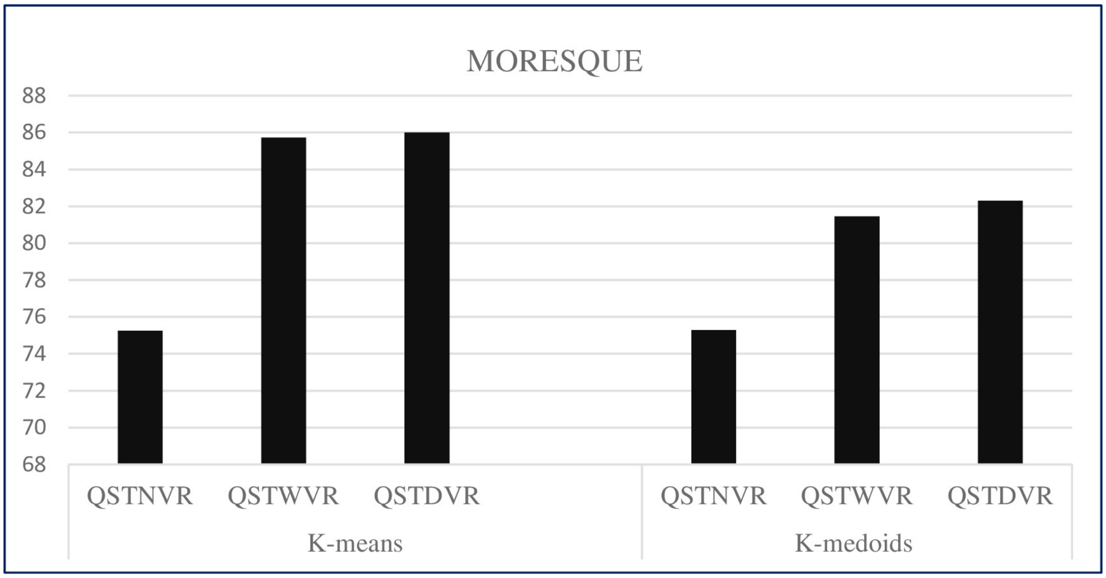

Fig 6. Performance of the three single-view models with k-means and k-medoids on the MORESQUE dataset.

https://doi.org/10.1371/journal.pone.0245264.g006

The stability of a clustering model can be defined as its ability to achieve the same results (performance) for all data types as shown in Figs 5 and 6. However, single-view clustering models are known for their instability, and no single clustering model is appropriate for all data types. The comparative results of the baseline clustering model on all datasets show that clustering performance varies from one dataset to another. The results in Tables 4 and 6 present the performance of each single-view model with k-means and k-medoids, and the presented results vary on different datasets.

In Subsection 5.3, the experimental results of the MMCC ensemble in Tables 8 and 9 demonstrate that a dynamic clustering model can be constructed by combining different candidate clustering solutions. These candidate clustering solutions combine high-quality and low-quality clusters to handle the dynamic nature of WSRC, where different queries indicate different search results and cluster numbers.

The experimental results in Tables 8 and 9 show that the proposed MMCC performs better than all single-view and representation clustering models. The performances of the single-view models with k-means and k-medoids are 86.36 and 82.03, respectively, on the ODP-239 dataset. Similarly, the performances of the single-view model with k-means and k-medoids are 86.01 and 84.69, respectively, on the MORESQUE dataset. However, the performances of the MMCC with k-means and k-medoids are 91.15 and 89.79, respectively, on the ODP-239 dataset. Integrating multiview methods with clustering methods is useful and improves the performance for all targeted datasets, as presented in Fig 7.

Applying multiple views leads to various clusters in different candidate clustering solutions. Using multiview multirepresentation simultaneously generates better clusters [27]. Hence, using multiple views that represent data in totally different representations with clustering methods and create a diverse set of candidate clustering solutions is logical. Multiview multirepresentation creates a diverse cluster and selects high-quality clusters from candidate clustering solutions, thereby leading to better clustering approach.

Additionally, the Wilcoxon rank statistical test [61] is used in this study to determine the significance of the results obtained by the MMCC model compared with the single-view models. A significance level α of 0.05 is employed for the test. Overall, the results show that the

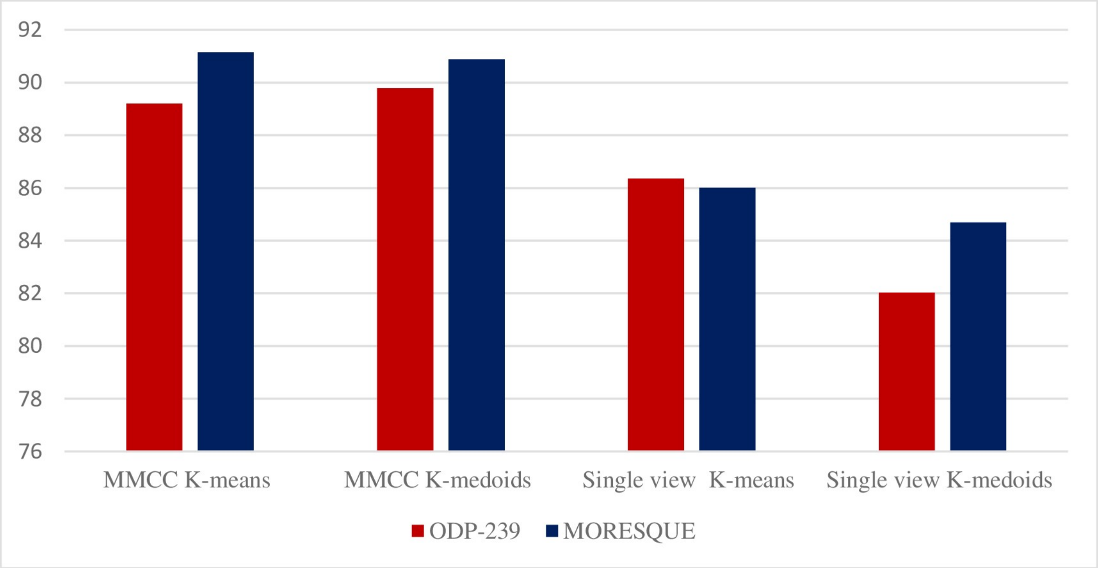

Fig 7. Overall performance of the single-view model and MMCC on the ODP-239 and MORESQUE datasets.

https://doi.org/10.1371/journal.pone.0245264.g007

MMCC is statistically more significant than the other compared clustering methods on all datasets with a p-value of 0.014.

In summary, this study enhances WSRC using the proposed MMCC that combines different views and representation methods to create different candidate clustering solutions. The proposed MMCC improves the performance of WSRC by achieving better results. Similarly, this study investigates all clustering methods' performance by conducting a statistical analysis, and the results show that the proposed MMCC is statistically significant over the compared methods.

7. Conclusion

This study presents a multiview multirepresentation approach with a CC ensemble method, which is also known as MMCC. This method uses multiple views to create a set of candidate solutions and selects high-quality clusters from the candidate solutions to form a superior clustering solution. This study contributes to improving the WRSC results by using three views and representations, namely, QSTNVR, QSTWVR, and QSTDVR with k-based clustering methods, namely, k-means and k-medoids, to generate the ensemble. These views and representations are used to create an initial set of candidate clustering solutions. This approach creates a diverse candidate clustering solution with combined high-quality and low-quality clusters. This study also proposes a radius-based cluster number estimation algorithm to handle the dynamic nature of WSRC. Two benchmark datasets are used to evaluate the proposed method and investigate the clustering techniques, among which MORESQUE and ODP-239 are extensively used. Moreover, the proposed MMCC outperforms the other methods in terms of precision, recall, and f-measure. The proposed method MMCC has a statically significant enhancement compared with other clustering methods based on the single-view model. Therefore, producing various sets of clusters from meaningful multiple views plays an essential role in WSRC clustering, but the existing clustering methods, including the proposed method, are vulnerable to poor cluster labeling and poor clustering quality due to overlapping problems. For future work, a new clustering method to address these limitations is proposed, and cost effectiveness and computational complexity can be considered. The current clustering methods attempt to cope with the drawbacks of the text data but cannot utilize the data with different representation methods to enhance clustering methods. Hence, better results in clustering methods could be achieved by employing data representation. On this basis, future researchers are encouraged to exploit features in enhancing text clustering methods from different views.

Author Contributions

Data curation: Ali Sabah.

Formal analysis: Ali Sabah.

Funding acquisition: Sabrina Tiun.

Methodology: Ali Sabah, Adil Yaseen Taha.

Supervision: Sabrina Tiun, Nor Samsiah Sani, Masri Ayob.

Writing – original draft: Ali Sabah.

Writing – review & editing: Sabrina Tiun.

References

1. Wahid A, Gao X, Andreae P. Multi-objective multiview clustering ensemble based on evolutionary approach. In2015 IEEE Congress on Evolutionary Computation (CEC) 2015 May 25 (pp. 1696–1703). IEEE.
2. Fraj M, Hajkacem MA, Essoussi N. Ensemble method for multi-view text clustering. InInternational Conference on Computational Collective Intelligence 2019 Sep 4 (pp. 219–231). Springer, Cham.
3. Abdulameer AS, Tiun S, Sani NS, Ayob M, Taha AY. Enhanced clustering models with wiki-based k-nearest neighbors-based representation for web search result clustering. Journal of King Saud University-Computer and Information Sciences. 2020 Feb 22.
4. Acharya S, Saha S, Moreno JG, Dias G. Multi-objective search results clustering. InProceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers 2014 Aug (pp. 99–108).
5. Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence. 2011 May; 25(03):337–72.
6. Yang Y, Wang H. Multi-view clustering: A survey. Big Data Mining and Analytics. 2018 Apr 12; 1(2):83–107.
7. Alam M, Sadaf K. A review on clustering of web search result. InAdvances in Computing and Information Technology 2013 (pp. 153–159). Springer, Berlin, Heidelberg.
8. Sadeghian AH, Nezamabadi-pour H. Document clustering using gravitational ensemble clustering. In2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP) 2015 Mar 3 (pp. 240–245). IEEE.
9. Zhou Z. H. Ensemble learning. Encyclopedia of biometrics. (2015), 411–416.
10. Abu-Jamous B, Fa R, Roberts DJ, Nandi AK. Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery. PLoS One. 2013 Feb 11; 8(2):e56432. https://doi.org/10.1371/journal.pone.0056432 PMID: 23409186
11. Feng Z, Bao J, Liu K. A New Efficient Text Clustering Ensemble Algorithm Based on Semantic Sequences. InInternational Conference in Swarm Intelligence 2013 Jun 12 (pp. 183–190). Springer, Berlin, Heidelberg.
12. Wang W. An improved non-negative matrix factorization algorithm for combining multiple clusterings. In2010 International Conference on Machine Vision and Human-machine Interface 2010 Apr 24 (pp. 604–607). IEEE.
13. Abu-Jamous B, Fa R, Nandi AK, Roberts DJ. Binarization of consensus partition matrix for ensemble clustering. In2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO) 2012 Aug 27 (pp. 2193–2197). IEEE.
14. Hussain SF, Mushtaq M, Halim Z. Multi-view document clustering via ensemble method. Journal of Intelligent Information Systems. 2014 Aug 1; 43(1):81–99.

15. Boongoen T, Iam-On N. Cluster ensembles: A survey of approaches with recent extensions and applications. Computer Science Review. 2018 May 1; 28:1–25.

16. Yin J, Wang J. A dirichlet multinomial mixture model-based approach for short text clustering. InProceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014 Aug 24 (pp. 233–242).

17. Wang Y, Zhang W, Wu L, Lin X, Fang M, Pan S. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. arXiv preprint arXiv:1608.05560. 2016 Aug 19.

18. Cao X, Zhang C, Fu H, Liu S, Zhang H. Diversity-induced multi-view subspace clustering. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 586–594).

19. Liu J, Wang C, Gao J, Han J. Multi-view clustering via joint nonnegative matrix factorization. InProceedings of the 2013 SIAM International Conference on Data Mining 2013 May 2 (pp. 252–260). Society for Industrial and Applied Mathematics.

20. Jiang Y, Liu J, Li Z, Lu H. Collaborative PLSA for multi-view clustering. InProceedings of the 21st International Conference on Pattern Recognition (ICPR2012) 2012 Nov 11 (pp. 2997–3000). IEEE.

21. Petkos G, Papadopoulos S, Kompatsiaris Y. Social event detection using multimodal clustering and integrating supervisory signals. InProceedings of the 2nd ACM International Conference on Multimedia Retrieval 2012 Jun 5 (pp. 1–8).

22. Samangooei S, Hare J, Dupplaw D, Niranjan M, Gibbins N, Lewis PH, et al. Social event detection via sparse multi-modal feature selection and incremental density based clustering.

23. Petkos G, Papadopoulos S, Schinas E, Kompatsiaris Y. Graph-based multimodal clustering for social event detection in large collections of images. InInternational Conference on Multimedia Modeling 2014 Jan 6 (pp. 146–158). Springer, Cham.

24. Sun J, Lu J, Xu T, Bi J. Multi-view sparse co-clustering via proximal alternating linearized minimization. InInternational Conference on Machine Learning 2015 Jun 1 (pp. 757–766).

25. Sun J, Bi J, Kranzler HR. Multi-view biclustering for genotype-phenotype association studies of complex diseases. In2013 IEEE International Conference on Bioinformatics and Biomedicine 2013 Dec 18 (pp.

26. Chao G, Sun J, Lu J, Wang AL, Langleben DD, Li CS, et al. Multi-view cluster analysis with incomplete data to understand treatment effects. Information Sciences. 2019 Aug 1; 494:278–93. https://doi.org/10.1016/j.ins.2019.04.039 PMID: 32863420

27. Bickel S, Scheffer T. Multi-view clustering. InICDM 2004 Nov 1 (Vol. 4, No. 2004, pp. 19–26).

28. Carpineto C, Romano G. Optimal meta search results clustering. InProceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval 2010 Jul 19 (pp. 170–177).

29. Saha S, Mitra S, Kramer S. Exploring multiobjective optimization for multiview clustering. ACM Transactions on Knowledge Discovery from Data (TKDD). 2018 May 23; 12(4):1–30.

30. Moreno JG, Dias G. Easy web search results clustering: When baselines can reach state-of-the-art algorithms. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers 2014 Apr (pp. 1–5).

31. Moreno JG, Dias G, Cleuziou G. Query log driven web search results clustering. InProceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval 2014 Jul 3 (pp. 777–786).

32. Navigli R, Crisafulli G. Inducing word senses to improve web search result clustering. InProceedings of the 2010 conference on empirical methods in natural language processing 2010 Oct (pp. 116–126).

33. Law MH, Topchy AP, Jain AK. Multiobjective data clustering. InProceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. 2004 Jun 27 (Vol. 2, pp. II-II). IEEE.

34. Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research. 2002; 3(Dec):583–617.

35. Dai H, Sheng W. A Multi-objective Clustering Ensemble Algorithm with Automatic k-Determination. In2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) 2019 Apr 12 (pp. 333–337). IEEE.

36. Wei C, Luo S, Ma X, Ren H, Zhang J, Pan L. Locally embedding autoencoders: a semi-supervised manifold learning approach of document representation. PloS one. 2016 Jan 19; 11(1):e0146672. https://doi.org/10.1371/journal.pone.0146672 PMID: 26784692

37. Mazarura J, De Waal A. A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech) 2016 (pp. 1–6). IEEE.

38. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003; 3 (Jan):993–1022.

39. Hui S, Dechao Z. A weighted topical document embedding based clustering method for news text. In2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference 2016 May 20 (pp. 1060–1065). IEEE.

40. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data. 2019 May 10; 6(1):1–9. https://doi.org/10.1038/s41597-018-0005-2 PMID: 30647409

41. Ibrahim R, Zeebaree S, Jacksi K. Survey on Semantic Similarity Based on Document Clustering. Adv. Sci. Technol. Eng. Syst. J. 2019; 4(5):115–22.

42. Luhn HP. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development. 1957 Oct; 1(4):309–17.

43. Sangaiah AK, Fakhry AE, Abdel-Basset M, El-henawy I. Arabic text clustering using improved clustering algorithms with dimensionality reduction. Cluster Computing. 2019 Mar 1; 22(2):4535–49.

44. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems 2013 (pp. 3111–3119).

45. Tiun S, Nor NF, Jalaludin A, Rahman AN. Word Embedding for Small and Domain-specific Malay Corpus. InComputational Science and Technology 2020 (pp. 435–443). Springer, Singapore. https://doi.org/10.1371/journal.pone.0242899 PMID: 33320858

46. Abd MT, Mohd M. A comparative study of word representation methods with conditional random fields and maximum entropy markov for bio-named entity recognition. Malaysian Journal of Computer Science. 2018 Dec 28:15–30.

47. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al's negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722. 2014 Feb 15.

48. Blei DM, Lafferty JD. Topic models. Text mining: classification, clustering, and applications. 2009 Jun 15; 10(71):34.

49. Rahman MA, Islam MZ. A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowledge-Based Systems. 2014 Nov 1; 71:345–65.

50. Hansen LK, Salamon P. Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence. 1990 Oct; 12(10):993–1001.

51. Tan AC, Gilbert D. An empirical comparison of supervised machine learning techniques in bioinformatics. In Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19 (pp. 219–222). Australian Computer Society.

52. Mousavi M, Bakar AA, Vakilian M. Data stream clustering algorithms: A review. Int J Adv Soft Comput Appl. 2015 Nov 2; 7(3):13.

53. Abdulameer AS, Saad S, Zakaria LQ. Trend detection in the Arabic social media using voting combination. Journal of Theoretical and Applied Information Technology. 2015 Nov 1; 81(3):621.

54. Teng G, He C, Xiao J, He Y, Zhu B, Jiang X. Cluster ensemble framework based on the group method of data handling. Applied Soft Computing. 2016 Jun 1; 43:35–46.

55. Alam M, Sadaf K. Web search result clustering based on heuristic search and K-means. arXiv preprint arXiv:1508.02552. 2015 Aug 11.

56. Zaw MM, Mon EE. Web Document Clustering Using Cuckoo Search Clustering Algorithm Based On Gauss Distribution. Transactions on GIGAKU: Scope and Policy. 2012 Jul:10.

57. Larsen B, Aone C. Fast and effective text mining using linear-time document clustering. InProceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining 1999 Aug 1 (pp. 16–22).

58. Durairaj M, Vijitha C. Educational data mining for prediction of student performance using clustering algorithms. International Journal of Computer Science and Information Technologies. 2014; 5(4):5987–91.

59. Karaa WB, Ashour AS, Sassi DB, Roy P, Kausar N, et al. Medline text mining: an enhancement genetic algorithm based approach for document clustering. InApplications of Intelligent Optimization in Biology and Medicine 2016 (pp. 267–287). Springer, Cham.

60. Mojahed A., & de la Iglesia B. (2017). An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach. Knowledge and Information Systems, 50(1), 27–52.

61. Marx A, Backes C, Meese E, Lenhof HP, Keller A. EDISON-WMW: exact dynamic programing solution of the Wilcoxon–Mann–Whitney test. Genomics, proteomics & bioinformatics. 2016 Feb 1; 14(1):55–61. https://doi.org/10.1016/j.gpb.2015.11.004 PMID: 26829645