The Journal of Supercomputing
https://doi.org/10.1007/s11227-021-03618-6

Sampling-based visual assessment computing techniques for an efficient social data clustering

M. Suleman Basha[1] · **S. K. Mouleeswaran**[1] · **K. Rajendra Prasad**[2]

Accepted: 2 January 2021

Abstract

Visual methods were used for pre-cluster assessment and useful cluster partitions. Existing visual methods, such as visual assessment tendency (VAT), spectral VAT (SpecVAT), cosine-based VAT (cVAT), and multi-viewpoints cosine-based similarity VAT (MVS-VAT), effectively assess the knowledge about the number of clusters or cluster tendency. Tweets data partitioning is underlying the problem of social data clustering. Cosine-based visual methods succeeded widely in text data clustering. Thus, cVAT and MVS-VAT are the best suited methods for the derivation of social data clusters. However, MVS-VAT is facing the problem of scalability issues in terms of computational time and memory allocation. Therefore, this paper presents the sampling-based MVS-VAT computing technique to overcome the scalability problem in social data clustering to select sample inter-cluster viewpoints. Standard health keywords and benchmarked TREC2017 and TREC2018 health keywords are taken to extract health tweets in the experiment for illustrating the performance comparison between existing and proposed visual methods.

Keywords Cluster tendency · Social data clustering · Scalability · Visual methods · Feature extraction

✉ M. Suleman Basha
suleman.ndl@gmail.com

S. K. Mouleeswaran
mouleeswaran-cse@dsu.edu.in

K. Rajendra Prasad
krprgm@gmail.com

[1] Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore, India

[2] Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, India

Published online: 12 January 2021

1 Introduction

Text clustering [1] derives the clusters based on similarity features of text documents. It is an emerging problem in many applications, such as social data clustering [2], information retrieval systems, media monitoring [3], feedback analysis, and opinion mining [4]. This paper attempts the problem of social data clustering for sensitive health data analysis [5]. Twitter is an excellent resource for obtaining social data [6]; it facilitates social users for sharing knowledge on health. Health tweets are extracted and modeled for the feature extraction of tweets in the form of bag-of-words. Topic models [7], 'non-matrix factorization (NMF)' [8], 'latent Dirichlet allocation (LDA)' [9], 'latent semantic indexing (LSI)' [10], and 'probabilistic LSI (PLSI)' [11] are the most popular in extraction of tweets features. With the topic models, the tweets' features are extracted concerning topics instead of terms for avoiding the data sparsity problem [1] [12]. State-of-the-art visual methods, VAT [13], SpecVAT [14], cVAT [15], and MVS-VAT [16], performed excellently for the extraction of cluster tendency, i.e., it determines the pre-clusters for the documents in visual form. The VAT was introduced to determine the clusters numbering with a count of extracted dark-black-colored squares in the respective images of visual methods, and the sample study of VAT is illustrated in Fig. 1 [13].

Dissimilarity features of documents are derived initially in dissimilarity matrix 'DM' and then find the re-ordered dissimilarity matrix (RDM [17]. Image of RDM is visualized, and it is known as a VAT image. VAT image showed the visual clusters in square-shaped dark-colored blocks. Each square-shaped dark-colored block of the VAT image indicates the separate cluster. Another method, SpecVAT, uses the spectral features of data objects in the dissimilarity features computation. Cosine-based similarity (or dissimilarity) values are computed in cVAT, in which the object's similarity measured with direction and magnitude of document vectors; thus, it is more accurate and shows the more quality of visual clusters. It uses the single-viewpoint approach. MVS-VAT uses the multi-viewpoints, and it conclusively determines the cluster tendency [18] for the set of tweets documents in a better way than the earlier stated visual methods. Finding the dissimilarity features of N tweets with (N-2) multi-viewpoints demands high computational time and more memory allocation. Here, dissimilarity was derived from any two tweets concerning remaining (N-2) viewpoints. Therefore, a sampling-based MVS-VAT method is proposed to address the scalability problems of social data clustering in terms of computational time and memory requirements, and its critical steps are shown in Fig. 2.

Figure 3 illustrates multi-viewpoints cosine-based similarity computation of the five tweets documents (here $N=5$), and corresponding documents are defined in projected space with five viewpoints, namely v1,v2, v3, v4, and v5. For example, similarity features between two documents (viewpoints v1 and v2) are computed concerning the other three multi-viewpoints (v3, v4, and v5), unlike a single viewpoint in traditional cosine metric.

The following cases are observed as follows:

$$DM = \begin{bmatrix} 0.00 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0.00 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0.00 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0.00 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0.00 \end{bmatrix}$$

(a) Distance (or Dissimilarity) Matrix of Dataset – DM; and its corresponding image (prior to VAT)

$$RDM = \begin{bmatrix} 0.00 & 0.12 & 0.59 & 0.73 & 0.78 \\ 0.12 & 0.00 & 0.59 & 0.71 & 0.74 \\ 0.59 & 0.55 & 0.00 & 0.19 & 0.19 \\ 0.73 & 0.71 & 0.19 & 0.00 & 0.16 \\ 0.78 & 0.74 & 0.19 & 0.16 & 0.00 \end{bmatrix}$$

(b) fter applying VAT, Re-ordered dissimilarity matrix (RDM), and its image 'I'

Fig. 1 Dissimilarity, re-ordered dissimilarity matrix, and visual images in VAT [13]

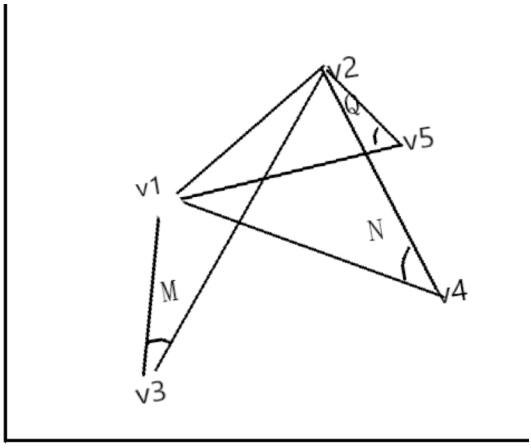

Fig. 2 Sampling-based view-points in similarity features computation

i. Cosine(v1,v2) with respect to v3 (with an angle M) is S1.
ii. Cosine(v1,v2) with respect to v4 (with an angle N) is S2.
iii. Cosine(v1,v2) with respect to v5 (with an angle Q) is S3.

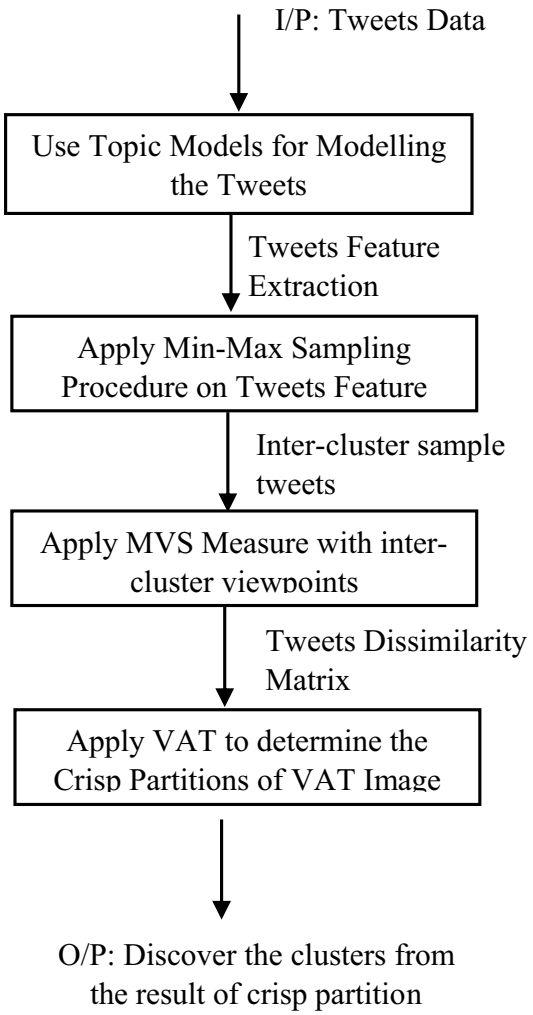

Fig. 3 Key steps of the proposed method

Highlights of the contributions for the described work of the paper are presented as follows:

1. The pre-cluster assessment is performed for the tweet dataset with the selection of the best sample viewpoints.
2. The sampling strategy is developed to select sample viewpoints rather than the selection of (N-2) viewpoints.
3. Cluster tendency is determined by developing the sampling-based visual method within minimum computational time.
4. The crisp partitions are derived for determining the cluster labels of health tweets.
5 The clusters are visualized with square-shaped black-colored blocks, which excellently determines the clusters of health tweets.

The crisp partitions have shown the aligned k-partitions of the visual image, and it is derived from the detection of obtained square-shaped blocks (dark color) along the diagonal in the visual image. The diagonal's square edginess showed the crisp partitions, and it is computed by finding the difference of pixel intensities between diagonal and non-diagonal square-shaped blocks.

The remaining sections are described as the following sequences. Related work is neatly presented in Sect. 2; the paper's proposed work is illustrated in Sect. 3. The essential part of the experimental work and its performance analysis is discussed in Sect. 4. Finally, the conclusion and scope of the work are described in Sect. 5.

2 Related work

Social data clustering involves two key steps: pre-clusters assessment and data partitions. Twitter [19] is a great social platform and provides social users an opportunity to share or exchange views through tweets form, for which social data clusters are important in related significant sectors. The clustering of social tweets is the current era of research in health domain applications. Health emergence needs to automate the social health data issues for finding high sentiment analysis [20]. For these reasons, many pre-cluster assessment methods are surveyed [14], and it was found that visual methods are recognized as the best choice for the smooth finding of cluster tendency. Bezdek et al. proposed the VAT [13], SpecVAT [14], improved VAT (iVAT) [21], and ClusiVAT [22] methods for the better assessment of clusters—this is observed in the respective state-of-the-art algorithms. VAT's basic approach is to initially find the dissimilarity features among the data objects using the Euclidean distance metric. Dissimilarity features are re-ordered in the resulting matrix 're-ordered dissimilarity matrix (RDM).' An algorithmic approach of VAT [13] is shown as follows (Algorithm 1).

```
Algorithm 1:VAT (int dissM[ ][ ],int n)
Step1:
   Let IV= { } ;JV={0,1,.....n-1}
Determine  max of dissM[ ] [ ], and its index
   cell is (i,j)
P(0)=i; IV={i},JV=JV-{IV};
Step2:
          for (s=1;s<n;s++)
          {
Find(i,j) from min {dissM[i][j], where
i€ IV, j є{JV}}
IV= {IV}U{j}; JV={JV}-{IV};
P(s)=j;
}
Step3:
/* Reordered Dissimilarity MatrixComutation*/
for(i=0;i<n;i++)
for(j=0;j<n;j++)
                RDM=dissM(P[i],P[j]);
Step 4:
    Display Image(RDM)
```

It has shown the visual clusters by displaying the re-ordered dissimilarity matrix. Each visual cluster represents a 'dark-black-colored block' in the visual image's diagonal (of VAT). The SpecVAT is applied to assess data objects with spectral features, which improves the clarity of visual clusters with spectral features and helps acquire adequate knowledge about cluster tendency. The problem of VAT is to produce the excellent cluster tendency assessment only for the limited size of dimensional datasets. The high-dimensional clusters assessment is well performed with the spectral features in SpecVAT. However, both VAT and SpecVAT are unable to handle complex datasets like path-shaped datasets. The iVAT handles this problem in the assessment of cluster tendency for path-shaped datasets. The critical approach of iVAT is to compute the path-based distances among the data objects; thus, it effectively works for path-shaped datasets. For big datasets, clustering with the sampling approach of ClusiVAT is developed to effectively address the big data cluster tendency problem. The limitation of ClusiVAT is less efficient for the text data clustering problem because text data clustering has come under the non-compact separated (non-CS) data. The ClusiVAT perfectly works for CS data rather than non-CS data.

Tweets are denoted as the text documents and initially modeled with topic models for deriving the features of tweet documents concerning topics instead of terms. Documents features versus terms are facing the data sparsity problem due to many specific terms that appeared in the tweet's documents. Thus, the feature vector for

the documents versus topics is the better choice for the clustering problem, and it is solved with the derivation of topics features of tweets documents with topic modeling techniques. Tweets features are derived in terms of topics, and it has less data sparsity when compared to defining the tweets features in 'term frequency (TF) and inverse document frequency (IDF),' a combined phrase known as TF-IDF [23]. Bag-of-words [24] for the tweets are expressed concerning topics which is the best choice of representation in social data clustering due to massive size tweets.

Finding the similarity features based on cosine produces a fair assessment of cluster tendency in cVAT [25] for the set of text (or tweets) documents. In cVAT, dissimilarity features are obtained concerning a single viewpoint, i.e., origin. Recently, MVS-VAT is developed for the social healthcare data clustering that computes the dissimilarity features concerning (N-2) multi-viewpoints instead of a single viewpoint for the high quality of social data clusters, where N indicates the documents count of tweets. Determining the clustering of tweet documents with (N-2) viewpoint demands high computational time and memory allocations. With reducing the complexities, sample viewpoints are needed instead of (N-2) in the pre-clusters assessment and finding the complete clustering results.

3 Proposed work

The proposed work aims to derive the social data clustering results with an extended approach—sampling-based viewpoints visual method. This work attempts to address the problem of MVS-VAT with the selection of sample viewpoints from the inter-cluster regions, which is presented in Algorithm 2; it is known as sample viewpoints cosine-based similarity VAT (SVPCS-VAT).

Algorithm 2

Input : N – Total Number of Tweets
Tweets Dataset {T1,T2,… Tn}
Output : Cluster Tendency 'k',
Tweets Data Clusters 'C'
Method :

3.1 Methods

Step 1: Extract tweets feature.

Model the tweets and extract the tweets features with respect to topics using topic models for the tweets $\{T_1, T_2,\ldots T_N\}$. Features of tweets are $\{F_1,F_2,\ldots F_N\}$.

Step 2: Find the centroid for initial cluster.

Select random number 'r' of {1, 2, …, N}. Find the distances between F_r and $\{F_1,F_2,\ldots F_N\}$ and choose the index based on maximum distance, and max_index = argmax $_{I \epsilon \{1,2,..N\}}$ {distance(F_r, F_I)} and max_dist = distance(F_r, F_I)); max_index shows the index of data object and has been selected as centroid.

Step 3: Assess the other centroids for topics.

Update the distances for the explored tweets.

For $i = 1$ to N.

$Dist_i = \min(max\_dist,$

$distance(F_{max\_index}, F_i)).$

Step 4: Update the other centroids.

Index of centroid is derived from argmax $_{I \epsilon \{1,2,..N\}}$ {$Dist_I$}, update max_index and max_dist, repeat Step 3 and Step 4 until obtain the topics centroids.

Step 5: Find the nearest tweets for the centroids.

Step 6: For $i = 1$ to N.

For $j = 1$ to N.

*Fix the sa*mple size and determine the sample viewpoints from the inter clusters of two tweet documents features (F_i,F_j). Compute the cosine similarity between these tweets with respective sample viewpoints (vp) of inter-clusters generated at earlier steps.

$$\mathrm{SVPCS(Di,Dj)} = \mathrm{avg}\left(\sum_{\mathrm{vpno=vp,Fj-vp}}^{\mathrm{npsample}} \mathrm{Sim(D1,D2)}\right)$$

$$\mathrm{Sim(D1,D2)} = \cos(\mathrm{Fi} - \mathrm{vp}, \mathrm{Fj} - \mathrm{vp})$$

Dissimilarity Feature(DM(i,j)) = (1- Normalize(SVPCS(Di,Dj),0,1).

Step 7: Apply VAT on DM and obtains the RDM.

Step 8: Display the image of RDM for the sample datasets for addressing unknown cluster tendency.

Step 9: Find the crisp partitions. With this cluster, labels of tweets are generated for discovering the tweets data clustering results 'C.'

Modeling the tweets with topic models and extracting bag-of-words features concerning topics are illustrated in Step 1. Bag-of-words are the feature vector that

consists of individual words. These bag-of-words are in the form of a word embedding model with their frequency counts of each tweet document. The topic model generates the topics-based bag-of-words instead of word embedding analysis in our proposed work. Pre-estimations of inter-clusters centroids are computed in Step 2 to Step 4—euclidean distance used for measuring the distances among the data objects for selection of centroids. The nearest cluster centroids of other tweet objects are defined in Step 5. Sample viewpoints are selected from inter-clusters estimations of earlier steps and used in cosine similarity computation of any two objects, which is explained in Step 6. Dissimilarity is computed with the subtraction of normalized similarity value from 1, and these values are stored in the matrix DM. Tweet objects are re-ordered based on DM, and its results are stored in RDM.

The values in DM represent the distances among the data objects; for example, the value of (i,j) location in DM denotes the distance between the ith and jth objects. Initially, data object 'i' is selected based on the maximum distance of DM for the location of (i, j). According to the minimum spanning tree (MST) cuts for the data objects, the indices of data objects are changed; the dissimilarity matrix's re-ordering is performed according to MST cut indices of data objects.

Step 7 describes the steps for a finding of DM and RDM. Step 8 shows the procedural steps for obtaining SVPCS-VAT image for presenting the output of cluster tendency. Step 9 shows the crisp partitions generations and tweets data clustering results. The crisp partitions are derived with the finding of squareness properties of appeared black-colored blocks along the diagonal of VAT image. Dark-colored blocks' squareness property is derived from the difference of pixel values between dark-colored blocks and non-dark-colored blocks.

Experimental details and performance study of visual methods for accessing the clustering tendency and tweets data clustering results are presented in the next section.

4 Experimental work and performance analysis

Tweets are extracted based on standard health keywords [26] and benchmarked TREC2017 [27] and TREC2018 [28] keywords. The tweets datasets for the experimental study are presented in Table 1, and sample tweets are given in Table 2.

Sample results of cluster tendency for the TREC2017 (2 topics), 15 topics, and TREC2018 (6 topics) are shown in Figs. 4, 5, and 6, respectively. Excellent clarity of the visual image has appeared with SVPCS-VAT method. The clustering tendency is determined by counting the total number of grey-shaded/dark-black-colored blocks. For the mentioned sample topics in the following figure, only SVPCS-VAT produces the quality of visual dark-colored blocks, which helps for the best assessment of cluster tendency. Therefore, the proposed visual SVPCS-VAT can access the high quality of social data clusters than other visual methods. Visual images have appeared with dark- or grey-colored blocks forming along the diagonal, and the numbers of square-shaped dark-colored blocks in Figs. 4, 5, and 6 are 2, 15, and 6, respectively. The count of blocks denoted the value of cluster tendency.

Table 1 Details of social data

S. no.	Dataset—nT (nT denotes the 'n' topics)	Used keywords during dataset creation	Size of the data in multiples of k (1000)
1	2 T	appendix, bone_density	80 k
2	3 T	appendix, bone_density, Brain Tumor	120 k
3	4 T	appendix, bone_density, Brain Tumor, common_cold	160 k
4	5 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes	200 k
5	6 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia	240 k
6	7 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones	280 k
7	8 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies	320 k
8	9 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies, Seizures	360 k
9	10 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies, Seizures, skin cancer	400 k
10	11 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies, Seizures, skin cancer, jaundice	440 k
11	12 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies, Seizures, skin cancer, jaundice, menopause	480 k
12	13 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies, Seizures, skin cancer, jaundice, menopause, obesity	520 k
13	14 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies, Seizures, skin cancer, jaundice, menopause, obesity, pneumonia	560 k
14	15 T	appendix, bone_density, Brain Tumor, common_cold, Diabetes, insomnia, kidney_stones, Rabies, Seizures, skin cancer, jaundice, menopause, obesity, pneumonia, swine flu	600 k
TREC 2017			
1	2 T	Constipation, diarrhea	80 k
2	3 T	Constipation, diarrhea, ebola	120 k

Table 1 (continued)

S. no.	Dataset—nT (nT denotes the 'n' topics)	Used keywords during dataset creation	Size of the data in multiples of k (1000)
TREC 2018			
1	2 T	Liposarcoma, Meningioma	80 k
2	3 T	Liposarcoma, Meningioma, Breast cancer	120 k
3	4 T	Liposarcoma, Meningioma, Breast cancer, Melanoma	160 k
4	5 T	Liposarcoma, Meningioma, Breast cancer, Melanoma, Gastrointestinal stromal tumor	200 k
5	6 T	Liposarcoma, Meningioma, Breast cancer, Melanoma, Gastrointestinal stromal tumor, carcinoma	240 k

Table 2 Sample tweets

Keyword	Sample tweet
Appendix	I was once turned away from the ER with a burst **appendix b**ecause I was trans, and that was in a liberal city in a country with free healthcare where non-discrimination laws protected my right to equal access
Bone density	Why does estrogen protect women from cardiovascular diseases, increases their **bone density**, and strengthens their skin and hair while testosterone cases balding and BPH? This isnt fair
Brain tumor	Two years ago, Penelope was diagnosed with a **brain tumor**. The tumor and surgery left her with some physical limitations. Every day, Penelope is working to get stronger and walk more
Common cold	something I hadn't really thought of through this pandemic is that my next non-COVID19 illness (seasonal flu, **common cold**) is going to be such an existential horror

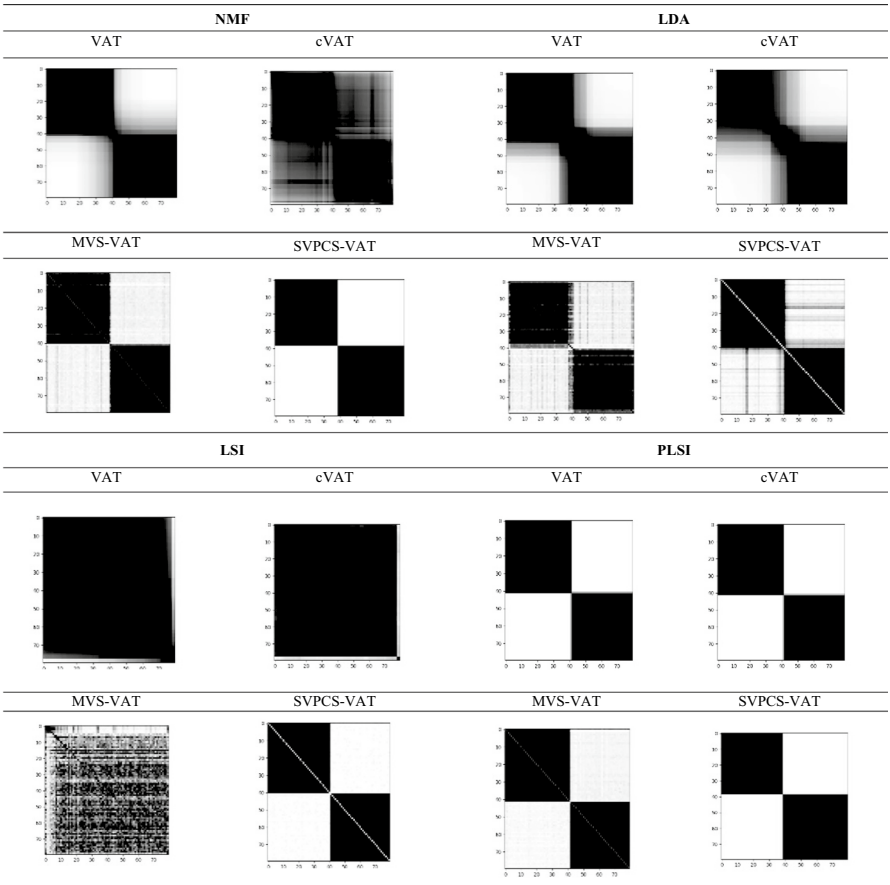

Fig. 4 Cluster tendency assessment for visual methods for TREC-2017 (2 keywords)

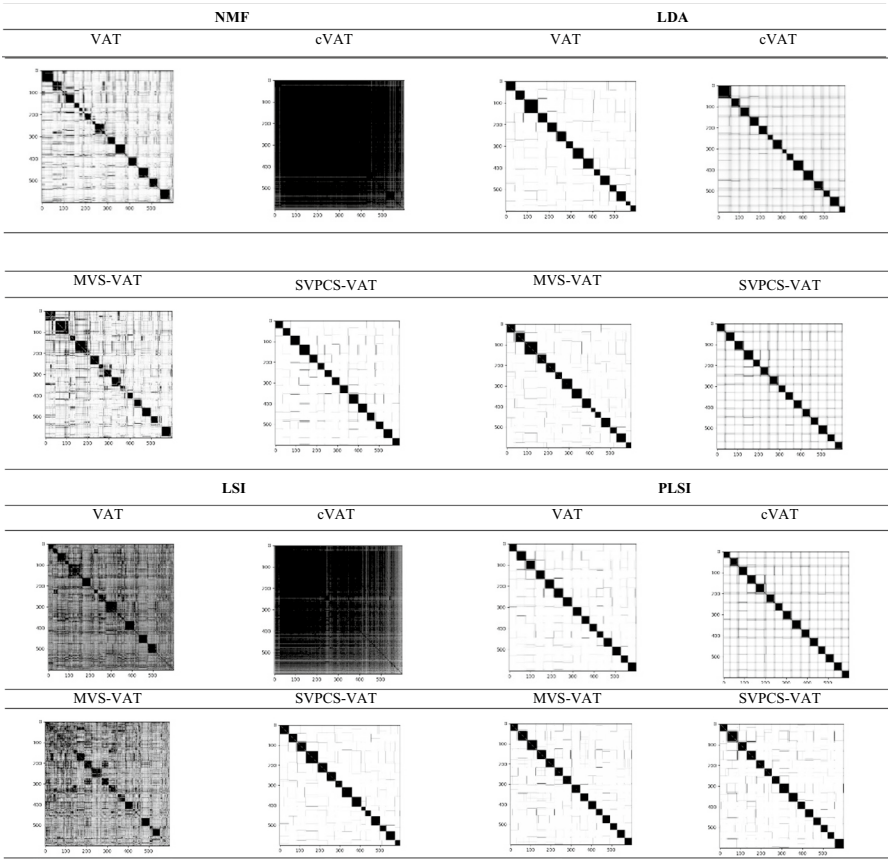

Fig. 5 Cluster tendency assessment for visual methods (for 15 topics)

The crisp partitions are extracted along with the finding of the edginess boundaries of dark-colored blocks along the diagonal. The crisp partitions are derived in Eq. (1), and parameters details are given in [29]. The crisp partitions illustration for the 3 topics of TREC2017 is shown in Fig. 7

$$f(U,D) = \left(\frac{\sum_{i=1}^{k} \sum_{s\in i,\tau no\tau\in i} d_{st}^{*}}{\sum_{i=1}^{k} (n-n_i)n_i} \right) - \left(\frac{\sum_{i=1}^{k} \sum_{s,\tau\in i,s\neq\tau} d_{st}^{*}}{\sum_{i=1}^{k} (n_i^2-n_i)} \right). \tag{1}$$

Inter-cluster viewpoints are selected with the proposed method, and they are used to find the dissimilarity features of tweets documents, in which good assessment of cluster tendency is achieved than other visual methods mentioned in Tables 3, 4, 5, 6, and 7. Table 8 presents the goodness of visual images for the visual methods. The best value of goodness represents the best assessment of clusters.

Performance evaluation conducted with five measures, namely 'cluster accuracy—CA' [30], 'normalized mutual information—NMI' [25], 'Precision—P,'

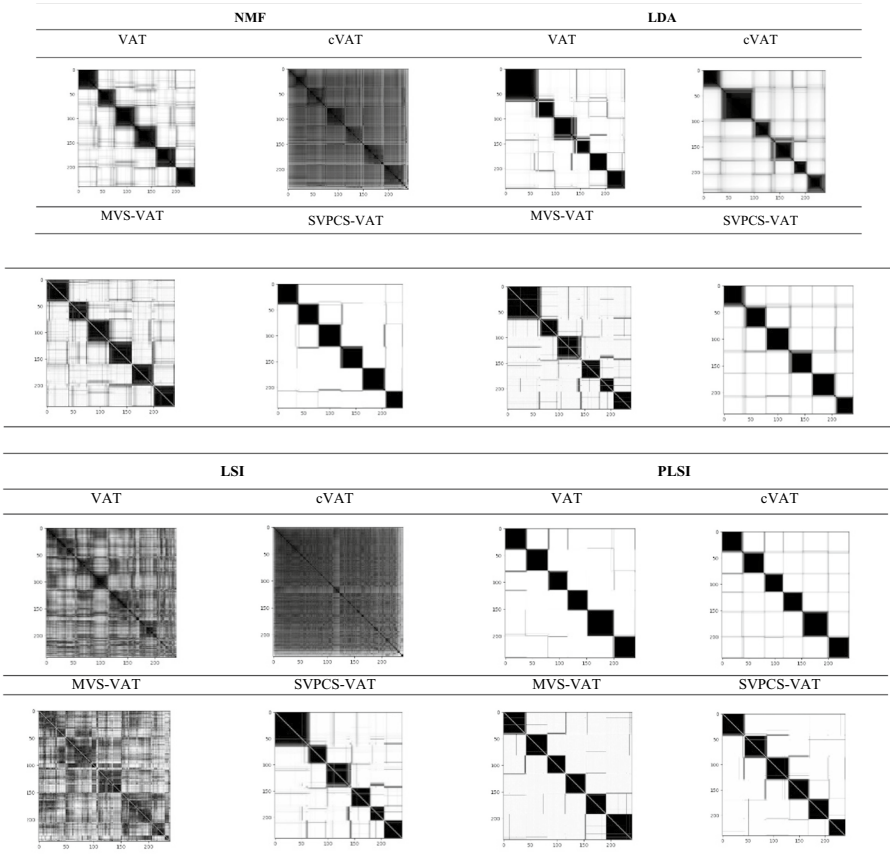

Fig. 6 Cluster tendency assessment for visual methods (TREC2018—6 topics)

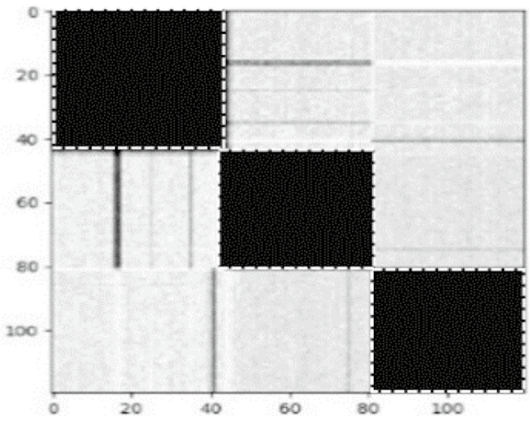

Fig. 7 Crisp partitions of SVPCS-VAT-NMF image (TREC2017—3 topics)

Table 3 Cluster accuracy (CA) for the visual methods

Number of topics for the dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT
2 T	0.800	0.800	0.525	0.852	0.575	0.575	0.575	0.575	0.850	0.675	0.500	0.858	0.525	0.525	0.500	0.535
3 T	1.000	1.000	1.000	1.000	0.583	0.583	0.542	0.598	0.775	0.633	0.650	0.785	0.375	0.442	0.375	0.452
4 T	1.000	1.000	0.988	1.000	0.413	0.419	0.419	0.425	0.631	0.706	0.556	0.769	0.356	0.350	0.344	0.381
5 T	0.765	0.765	0.800	0.830	0.295	0.300	0.290	0.330	0.600	0.445	0.425	0.612	0.315	0.340	0.325	0.335
6 T	0.817	0.817	0.829	0.835	0.308	0.292	0.292	0.321	0.575	0.396	0.450	0.581	0.258	0.254	0.254	0.271
7 T	0.593	0.593	0.736	0.742	0.271	0.250	0.264	0.278	0.454	0.382	0.382	0.532	0.257	0.275	0.250	0.289
8 T	0.850	0.850	0.906	0.906	0.225	0.216	0.244	0.244	0.459	0.341	0.444	0.481	0.200	0.203	0.209	0.209
9 T	0.742	0.742	0.667	0.758	0.231	0.217	0.239	0.239	0.453	0.300	0.500	0.500	0.194	0.194	0.186	0.210
10 T	0.653	0.653	0.715	0.725	0.215	0.185	0.193	0.221	0.443	0.338	0.550	0.562	0.220	0.215	0.208	0.225
11 T	0.668	0.668	0.711	0.721	0.205	0.198	0.186	0.214	0.543	0.323	0.500	0.552	0.205	0.202	0.202	0.214
12 T	0.696	0.696	0.619	0.710	0.213	0.192	0.177	0.221	0.519	0.290	0.515	0.540	0.194	0.204	0.190	0.218
13 T	0.562	0.562	0.577	0.610	0.187	0.169	0.169	0.195	0.448	0.283	0.442	0.481	0.175	0.177	0.158	0.185
14 T	0.495	0.495	0.523	0.527	0.173	0.171	0.152	0.182	0.463	0.254	0.488	0.492	0.180	0.195	0.166	0.211
15 T	0.477	0.477	0.450	0.495	0.172	0.158	0.163	0.179	0.513	0.242	0.437	0.525	0.163	0.165	0.168	0.176

Table 4 Normalized mutual information (NMI) for the visual methods

Number of topics for the dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT
2 T	0.278	0.278	0.002	0.292	0.016	0.016	0.016	0.016	0.390	0.090	0.000	0.421	0.002	0.002	0.000	0.007
3 T	1.000	1.000	1.000	1.000	0.237	0.237	0.204	0.248	0.482	0.324	0.203	0.497	0.009	0.032	0.009	0.033
4 T	1.000	1.000	0.958	1.000	0.084	0.090	0.091	0.099	0.428	0.422	0.324	0.549	0.059	0.053	0.045	0.075
5 T	0.673	0.673	0.671	0.711	0.070	0.053	0.061	0.077	0.377	0.210	0.240	0.385	0.059	0.063	0.060	0.065
6 T	0.767	0.767	0.792	0.812	0.100	0.070	0.070	0.121	0.383	0.240	0.342	0.398	0.074	0.061	0.049	0.081
7 T	0.547	0.547	0.667	0.687	0.116	0.101	0.093	0.125	0.333	0.235	0.280	0.364	0.058	0.075	0.058	0.082
8 T	0.749	0.749	0.832	0.832	0.093	0.063	0.072	0.098	0.349	0.232	0.295	0.358	0.062	0.062	0.070	0.070
9 T	0.627	0.627	0.610	0.639	0.095	0.081	0.084	0.101	0.376	0.197	0.394	0.394	0.070	0.060	0.064	0.075
10 T	0.636	0.636	0.675	0.682	0.096	0.085	0.079	0.114	0.413	0.227	0.471	0.489	0.109	0.103	0.095	0.112
11 T	0.625	0.625	0.666	0.684	0.105	0.077	0.077	0.117	0.468	0.241	0.454	0.471	0.100	0.107	0.108	0.118
12 T	0.641	0.641	0.630	0.641	0.135	0.115	0.101	0.148	0.494	0.224	0.455	0.521	0.098	0.107	0.101	0.110
13 T	0.518	0.518	0.538	0.558	0.118	0.105	0.099	0.128	0.454	0.233	0.418	0.469	0.099	0.108	0.088	0.118
14 T	0.463	0.463	0.502	0.503	0.110	0.109	0.098	0.124	0.488	0.217	0.462	0.495	0.124	0.117	0.105	0.132
15 T	0.498	0.498	0.486	0.498	0.115	0.095	0.098	0.134	0.491	0.207	0.446	0.511	0.105	0.116	0.113	0.118

Table 5 Precision (p) for the visual methods

Number of topics for the dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT
2 T	0.800	0.800	0.525	0.812	0.575	0.575	0.575	0.575	0.850	0.675	0.500	0.856	0.525	0.525	0.500	0.550
3 T	1.000	1.000	1.000	1.000	0.583	0.583	0.542	0.591	0.775	0.633	0.650	0.785	0.375	0.442	0.375	0.451
4 T	1.000	1.000	0.988	1.000	0.413	0.419	0.419	0.421	0.631	0.706	0.556	0.769	0.356	0.350	0.344	0.381
5 T	0.765	0.765	0.800	0.830	0.295	0.300	0.290	0.330	0.600	0.445	0.425	0.610	0.315	0.340	0.325	0.520
6 T	0.817	0.817	0.829	0.835	0.308	0.292	0.292	0.298	0.575	0.396	0.450	0.582	0.258	0.254	0.254	0.271
7 T	0.593	0.593	0.736	0.741	0.271	0.250	0.264	0.268	0.454	0.382	0.382	0.532	0.257	0.275	0.250	0.285
8 T	0.850	0.850	0.906	0.906	0.225	0.216	0.244	0.244	0.459	0.341	0.444	0.468	0.200	0.203	0.209	0.209
9 T	0.742	0.742	0.667	0.751	0.231	0.217	0.239	0.239	0.453	0.300	0.500	0.512	0.194	0.194	0.186	0.210
10 T	0.653	0.653	0.715	0.719	0.215	0.185	0.193	0.195	0.443	0.338	0.550	0.562	0.220	0.215	0.208	0.225
11 T	0.668	0.668	0.711	0.721	0.205	0.198	0.186	0.210	0.543	0.323	0.500	0.512	0.205	0.202	0.202	0.210
12 T	0.696	0.696	0.619	0.705	0.213	0.192	0.177	0.216	0.519	0.290	0.515	0.540	0.194	0.204	0.190	0.214
13 T	0.562	0.562	0.577	0.583	0.187	0.169	0.169	0.191	0.448	0.283	0.442	0.452	0.175	0.177	0.158	0.200
14 T	0.495	0.495	0.523	0.527	0.173	0.171	0.152	0.178	0.463	0.254	0.488	0.498	0.180	0.195	0.166	0.214
15 T	0.477	0.477	0.450	0.481	0.172	0.158	0.163	0.178	0.513	0.242	0.437	0.525	0.163	0.165	0.168	0.175

Table 6 Recall (R) for the visual methods

Number of topics for the dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT
2 T	0.802	0.803	0.532	0.810	0.581	0.585	0.585	0.591	0.662	0.668	0.595	0.672	0.512	0.515	0.521	0.534
3 T	1.000	1.000	1.000	1.000	0.585	0.856	0.552	0.861	0.752	0.698	0.721	0.758	0.352	0.359	0.368	0.378
4 T	1.000	1.000	0.992	1.000	0.415	0.419	0.421	0.425	0.635	0.710	0.625	0.715	0.352	0.350	0.359	0.374
5 T	0.768	0.762	0.812	0.841	0.281	0.291	0.291	0.312	0.605	0.512	0.459	0.612	0.305	0.312	0.325	0.338
6 T	0.819	0.821	0.831	0.832	0.308	0.295	0.296	0.311	0.578	0.395	0.495	0.581	0.220	0.234	0.248	0.258
7 T	0.595	0.595	0.739	0.741	0.251	0.258	0.259	0.268	0.459	0.462	0.461	0.478	0.250	0.261	0.271	0.289
8 T	0.851	0.852	0.910	0.912	0.220	0.221	0.231	0.241	0.462	0.465	0.468	0.472	0.201	0.211	0.221	0.238
9 T	0.731	0.735	0.651	0.739	0.221	0.218	0.228	0.250	0.441	0.421	0.429	0.482	0.185	0.195	0.210	0.228
10 T	0.642	0.649	0.652	0.672	0.210	0.198	0.199	0.212	0.421	0.408	0.421	0.429	0.214	0.224	0.241	0.258
11 T	0.669	0.667	0.672	0.682	0.207	0.204	0.198	0.212	0.540	0.512	0.541	0.552	0.210	0.225	0.238	0.249
12 T	0.652	0.659	0.651	0.662	0.207	0.198	0.195	0.214	0.510	0.512	0.518	0.525	0.185	0.195	0.214	0.225
13 T	0.551	0.552	0.558	0.562	0.175	0.169	0.171	0.179	0.449	0.425	0.432	0.458	0.168	0.171	0.182	0.210
14 T	0.491	0.491	0.521	0.528	0.162	0.168	0.167	0.175	0.465	0.425	0.452	0.475	0.210	0.221	0.234	0.252
15 T	0.452	0.452	0.450	0.462	0.165	0.159	0.162	0.172	0.510	0.508	0.495	0.521	0.157	0.168	0.179	0.195

Table 7 F-Measure (F) for the visual methods

Number of topics for the dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT
2 T	0.803	0.805	0.536	0.810	0.575	0.579	0.571	0.581	0.652	0.653	0.658	0.682	0.521	0.532	0.541	0.558
3 T	1.000	1.000	1.000	1.000	0.579	0.821	0.579	0.531	0.721	0.701	0.708	0.728	0.420	0.358	0.361	0.378
4 T	1.000	1.000	0.995	1.000	0.412	0.421	0.423	0.426	0.621	0.614	0.624	0.635	0.250	0.347	0.365	0.389
5 T	0.765	0.769	0.815	0.821	0.280	0.285	0.289	0.295	0.602	0.558	0.528	0.610	0.310	0.324	0.347	0.368
6 T	0.821	0.821	0.826	0.832	0.310	0.299	0.298	0.315	0.555	0.458	0.501	0.568	0.214	0.234	0.257	0.287
7 T	0.598	0.598	0.742	0.745	0.252	0.261	0.262	0.268	0.441	0.448	0.449	0.452	0.250	0.261	0.278	0.289
8 T	0.856	0.854	0.912	0.918	0.225	0.226	0.231	0.238	0.451	0.462	0.468	0.475	0.210	0.224	0.238	0.249
9 T	0.735	0.736	0.659	0.741	0.228	0.221	0.231	0.235	0.442	0.448	0.452	0.468	0.175	0.185	0.195	0.212
10 T	0.648	0.651	0.658	0.678	0.215	0.210	0.214	0.221	0.418	0.425	0.432	0.457	0.214	0.221	0.234	0.245
11 T	0.671	0.678	0.679	0.689	0.205	0.214	0.216	0.224	0.525	0.529	0.536	0.558	0.215	0.221	0.234	0.251
12 T	0.653	0.658	0.661	0.668	0.205	0.204	0.208	0.212	0.507	0.512	0.524	0.538	0.198	0.210	0.221	0.242
13 T	0.558	0.559	0.559	0.568	0.198	0.181	0.185	0.191	0.432	0.438	0.442	0.452	0.178	0.185	0.192	0.214
14 T	0.498	0.499	0.512	0.521	0.165	0.171	0.017	0.176	0.415	0.428	0.432	0.448	0.207	0.214	0.235	0.248
15 T	0.458	0.459	0.458	0.471	0.164	0.162	0.163	0.169	0.508	0.512	0.528	0.539	0.168	0.174	0.189	0.210

Table 8 Goodness of visual methods

Number of topics for the dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT	VAT	cVAT	MVS-VAT	SVPCS-VAT
2 T	0.010	0.825	0.722	0.982	0.525	0.528	0.506	0.685	0.808	0.747	0.522	0.811	0.575	0.571	0.552	0.582
3 T	0.759	0.761	0.784	0.921	0.411	0.409	0.397	0.528	0.671	0.454	0.736	0.739	0.421	0.425	0.482	0.485
4 T	0.985	0.985	0.975	0.997	0.324	0.345	0.321	0.491	0.591	0.481	0.682	0.691	0.323	0.325	0.292	0.335
5 T	0.825	0.866	0.953	0.967	0.305	0.307	0.295	0.459	0.563	0.446	0.493	0.572	0.321	0.308	0.282	0.527
6 T	0.937	0.939	0.941	0.945	0.261	0.261	0.268	0.427	0.537	0.479	0.494	0.545	0.265	0.272	0.251	0.572
7 T	0.922	0.924	0.954	0.959	0.258	0.255	0.255	0.325	0.451	0.457	0.456	0.461	0.263	0.269	0.282	0.325
8 T	0.671	0.677	0.654	0.825	0.217	0.213	0.215	0.355	0.432	0.384	0.547	0.549	0.331	0.435	0.445	0.645
9 T	0.645	0.646	0.682	0.795	0.214	0.224	0.219	0.391	0.417	0.317	0.412	0.623	0.208	0.208	0.384	0.415
10 T	0.605	0.594	0.572	0.812	0.213	0.207	0.215	0.412	0.461	0.312	0.441	0.725	0.315	0.411	0.495	0.519
11 T	0.507	0.507	0.498	0.798	0.198	0.186	0.197	0.457	0.527	0.554	0.687	0.815	0.291	0.586	0.383	0.798
12 T	0.541	0.544	0.471	0.655	0.205	0.173	0.198	0.392	0.385	0.265	0.422	0.427	0.192	0.188	0.181	0.199
13 T	0.485	0.482	0.442	0.589	0.198	0.162	0.184	0.367	0.384	0.227	0.384	0.398	0.187	0.189	0.178	0.195
14 T	0.465	0.464	0.494	0.702	0.194	0.178	0.183	0.310	0.322	0.208	0.325	0.338	0.164	0.172	0.175	0.181
15 T	0.422	0.419	0.432	0.725	0.177	0.168	0.144	0.327	0.342	0.191	0.327	0.351	0.159	0.165	0.163	0.172

'Recall—R,' and 'F-Measure—F' [31]. It is observed that the proposed SVPCS-VAT scored the highest performance values compared to others under various topics of tweets datasets.

Social data clustering results are derived with visual methods, and these are evaluated with four different topic modelling techniques (NMF-Features, LDA-Features, LSI-Features, and PLSI-Features). Crisp partitions are derived for the clustering of mentioned tweets data. Based on the ground-truth labels of tweets and predicted values, finally, the confusion matrix values are computed. Tweets classification model correctly predicts the positive and negative classes are referred to as true positive and true negative, respectively. In contrast, the model incorrectly predicts the positive and negative classes; then, they are false positives and negatives. With the confusion matrix values, visual methods' performance is computed with the following measures: Precision—P, Recall—R, and F-Measure—F. In all the cases of topic models, i.e., NMF, LDA, LSI, and PLSI, it is observed that proposed SVPCS-VAT scored the best performance scores compared to VAT, cVAT, and MVS-VAT.

The proposed SVPCS-VAT uses a few viewpoints for instead of (N-2) viewpoints, unlike MVS-VAT. The particular sample viewpoints are from the inter-clusters only, and the sampling-based viewpoints procedure has taken less amount of computation time and memory allocation during the generation of tweets data clustering results. The experimental study is carried out on tweets dataset for the 2 topics to 15 topics. Figures 8 and 9 show the time and space analysis of visual methods, which illustrates that the proposed SVPCS-VAT requires less time and memory requirements; hence, the proposed method achieves a suitable scalability MVS-VAT method. Performance analysis of visual methods in the experiment illustrates that SVPCS-VAT is a useful visual method that effectively accesses cluster tendency and discovers the quality of clusters for the tweet's dataset.

The speedup of SVPCS-VAT is estimated based on calculating the quotient concerning the computational (wall) time of MVS-VAT. Figure 10 shows the speedup values of SVPCS-VAT with respect to MVS-VAT. It was observed that the fastness of SVPCS-VAT is much improved. Experiment is conducted for the sample data in our proposed work, and empirical analysis is performed for the topics (2 topics to 15 topics). With the overall empirical analysis, it is observed that our proposed methods SVPCS-VAT-NMF, SVPCS-VAT-LDA, SVPCS-VAT-LSI, and SVPCS-VAT- PLSI are time and space efficient when compared to MVS-VAT-NMF, MVS-VAT-LDA, MVS-VAT-LSI, and MVS-VAT-PLSI, respectively. These methods have also outperformed the others concerning CA, NMF, P, R, and F. Among the four proposed models, SVPCS-VAT-NMF performed as the best for the large tweets data clustering.

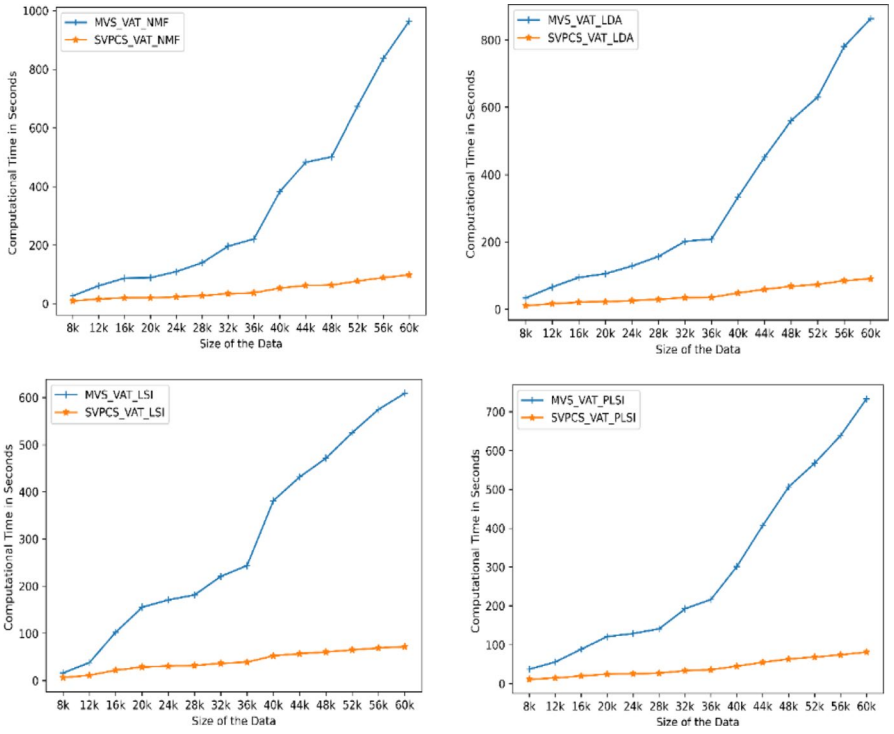

Fig. 8 Comparison of computational time for the visual methods—MVS-VAT and SVPCS-VAT

5 Conclusion and Future Scope

Accessing the clustering tendency is the crucial pre-clustering problem, which is useful for improving the quality of social data clusters. Health data clustering is an emerging need for society; thus, health-related tweets are extracted for finding the tendency of health data over social media (Twitter). Existing visual method MVS-VAT can assess the clustering tendency with (N-2) viewpoints to demand high computational time and memory allocation. The proposed sampling-based visual method, SVPCS-VAT, overcomes the complexity issues in social data clustering. Visual methods need to be extended with the parallel distributed techniques for handling the issues related to big social data clustering applications.

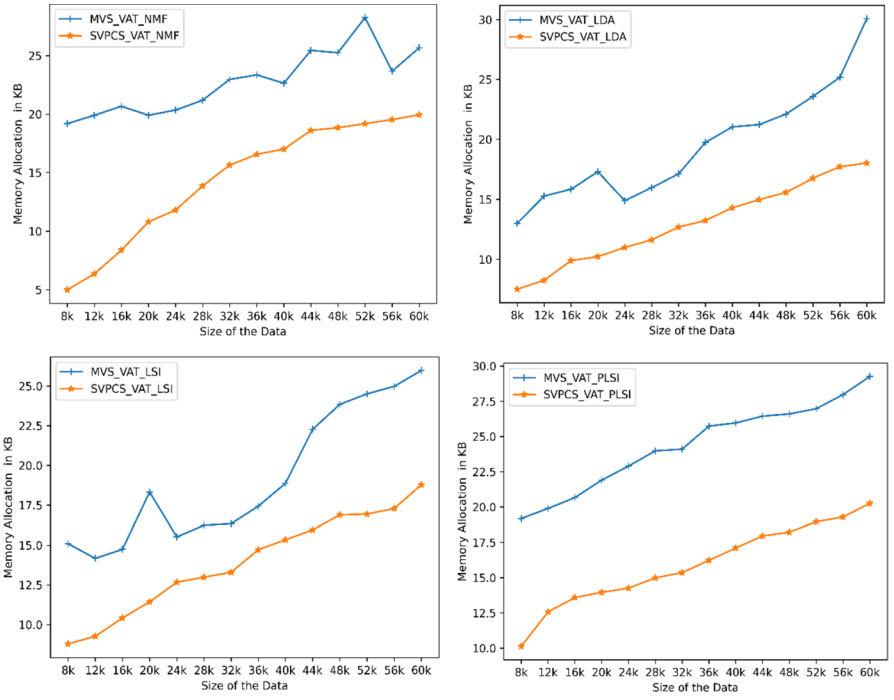

Fig. 9 Comparison of memory allocation for the visual methods—MVS-VAT and SVPCS-VAT

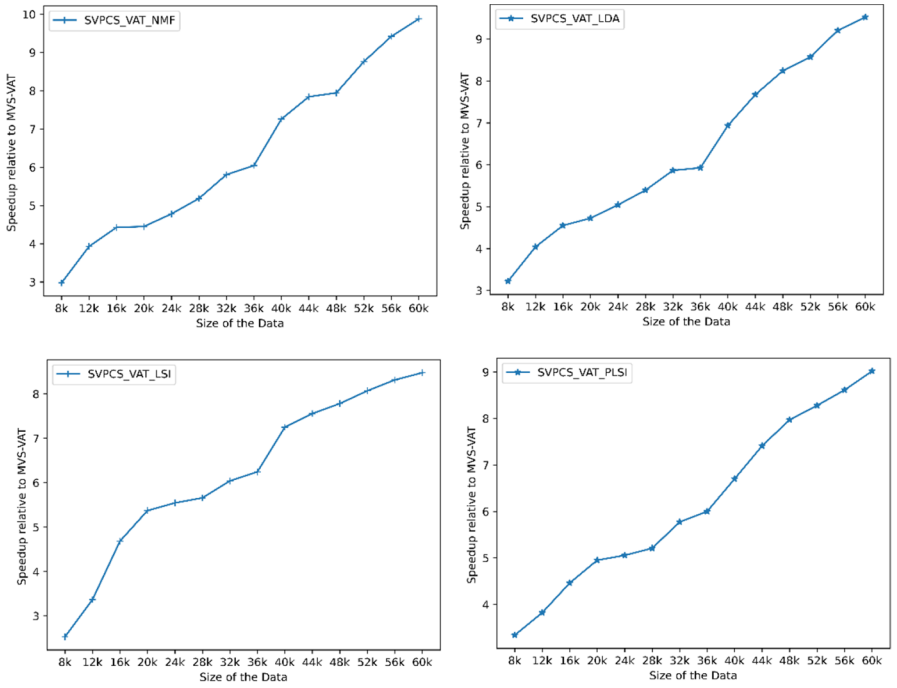

Fig.10 SVPCS-VAT Speedup relative to MVS-VAT

Acknowledgment This work is supported by the Science & Engineering Research Board (SERB), Department of Science and Technology, Government of India for the Research Grant of DST Project Number ECR/2016/001556.

References

1. Lin YS, Jiang JY, Lee SJ (2014) A similarity measure for text classification and clustering. IEEE Trans Knowledge Data Eng (2014)
2. Rui X, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678
3. Rajendra Prasad K, Suleman Basha M (2016) Improving the performance of speech clustering method. In: IEEE 10th International Conference on Intelligent Systems and Control (ISCO).
4. Wu X, Kumar V, Quinlan JR et al (2008) Top 10 algorithms in data mining, knowledge information system, vol 14. Springer, Heidelberg, pp 1–37.
5. Sik-Lanyi et al (2019) Accessibility testing of European health-related websites. Arab J Sci Eng 44:9171–9190
6. Ramathilagam S, Devi R, Kannan SR (2013) Extended fuzzy c-means: an analyzing data clustering problems. Cluster Comput
7. Feng Yi, Bo Jiang, Jianjun Wu (2020) Topic modeling for short texts via word embedding and document correlation. IEEE Access 8:30692–30705
8. Lee D, Seung H (2000) Algorithms for non-negative matrix factorization. Advances in neural information processing systems 13, NIPS 2000. Denver, CO, USA, pp 556–562
9. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
10. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

11. T Hofmann (1999) Probabilistic latent semantic indexing. SIGIR. ACM, New York, pp 50–57
12. Xu G, Meng Y, Chen Z, Qiu X, Wang C, Yao H (2019) Research on topic detection and tracking for online news texts. IEEE Access 7:58407–58418
13. Bezdek JC, Hathaway RJ (2002) VAT: a tool for visual assessment of (cluster) tendency. In: Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02, 2002, pp 2225–2230
14. Bezdek, James Leckie (2008) SpecVAT: enhanced visual cluster analysis. IEEE Int Conf Data Mining, ICDM
15. Rajendra Prasad K, Mohammed M, Noorullah RM (2019) Visual topic models for healthcare data clustering. Evolutionary Intelligence.
16. Rajendra Prasad K, Mohammed M, Noorullah RM (2019) Hybrid topic cluster models for social healthcare data. Int J Adv Comput Sci Appl 10(11):490–506.
17. Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah (2015) A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS 10(12):1–20
18. Suleman Basha M, Mouleeswaran SK, Rajendra Prasad K (2019) Cluster tendency methods for visualizing the data partitions. Int J Innovative Technol Explor Eng.
19. Vijeya Kaveri V, Maheswari V (2019) A framework for recommending health-related topics based on topic modeling in conversational data (Twitter). Cluster Computing.
20. Asghar MZ et al (2018) RIFT: a rule induction framework for twitter sentiment analysis. Arab J Sci Eng 43:857–877
21. Kumar D, Bezdek JC, Palaniswami M, Rajasegarar S, Leckie C, Havens TC (2016) A hybrid approach to clustering in big data. IEEE Trans Cybern 46(10):2372–2385
22. Kumar D, Palaniswami M, Rajasegarar S, Leckie C, Bezdek JC, Havens TC (2013) clusiVAT: A mixed visual/numerical clustering algorithm for big data. *2013* IEEE International Conference on Big Data, Silicon Valley, CA, 2013, pp 112–117.
23. Wuhan (2018) TF-IDF based feature words extraction and topic modeling for short text. In: ICMSS2018.
24. Wallach, Hanna M (2006) Topic modeling: beyond bag-of-words, ACM International Conference Proceeding Series, 2006
25. Alessia Amelio, Clara Pizzuti (2015) Is normalized mutual information a fair measure for comparing community detection methods?. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.
26. https://www.webmd.com/
27. https://trec.nist.gov/data/web2014.html
28. https://trec.nist.gov/data/microblog2015.html
29. Bodjanova S (2006) Crisp partitions Induced by a fuzzy set. In: Batagelj V, Bock HH, Ferligoj A., Žiberna A (eds) Data science and classification. Studies in classification, data analysis, and knowledge organization. Springer, Berlin (2006)
30. Pattanodom et al. (2016) Clustering data with the presence of missing values by ensemble approach. In: Second Asian Conference on Defense Technology.
31. Bhatnagar V, Majhi R, Jena PR (2018) Comparative performance evaluation of clustering algorithms for grouping manufacturing firms. Arab J Sci Eng 43:4071–4083