Query-based Topic Detection Using Concepts and Named Entities

Ilias Gialampoukidis[1], Dimitris Liparas[1], Stefanos Vrochidis[1], and Ioannis Kompatsiaris[1]

Abstract. In this paper, we present a framework for topic detection in news articles. The framework receives as input the results retrieved from a query-based search and clusters them by topic. To this end, the recently introduced “DBSCAN-Martingale” method for automatically estimating the number of topics and the well-established Latent Dirichlet Allocation topic modelling approach for the assignment of news articles into topics of interest, are utilized. Furthermore, the proposed query-based topic detection framework works on high-level textual features (such as concepts and named entities) that are extracted from news articles. Our topic detection approach is tackled as a text clustering task, without knowing the number of clusters and compares favorably to several text clustering approaches, in a public dataset of retrieved results, with respect to four representative queries.

1 INTRODUCTION

The need by both journalists and media monitoring companies to master large amounts of news articles produced on a daily basis, in order to identify and detect interesting topics and events, has highlighted the importance of the topic detection task. In general, topic detection aims at grouping together stories-documents that discuss about the same topic-event. Formally, a topic is defined in [1] as “*a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences*”. It is clarified [1] that the notion of “topic” is not general like “accidents” but is limited to a specific collection of related events of the type accident, such as “cable car crash”. We shall refer to topics as news clusters, or simply clusters.

The two main challenges involved in the topic detection problem are the following: one needs to (1) estimate the correct number of topics/news clusters and (2) assign the most similar news articles into clusters. In addition, the following assumptions must be made: Firstly, real data is highly noisy and the number of clusters is not known a priori. Secondly, there is a lower bound for the minimum number of documents per news cluster.

In this context, we present and describe the hybrid clustering framework for topic detection, which has been developed within the FP7 MULTISENSOR project[2]. For a given query-based search, the main idea is to efficiently cluster the retrieved results, without the need for a pre-specified number of topics. To this end, the framework, recently introduced in [2], combines automatic estimation of the number of clusters and assignment of news articles into topics of interest, on the results of a text query. The estimation of the number of clusters is done by the novel “DBSCAN-Martingale” method [2], which can deal with the aforementioned assumptions. All clusters are progressively extracted (by a density-based algorithm) by applying Doob’s martingale and then Latent Dirichlet Allocation is applied for the assignment of news articles to topics. Contrary to [2], the contribution of this paper is based on the fact that the overall framework relies on high-level textual features (concepts and named entities) that are extracted from the retrieved results of a textual query, and can assist any search engine.

The rest of the paper is organized as follows: Section 2 provides related work with respect to topic detection, news clustering and density-based clustering. In Section 3, our framework for topic detection is presented and described. Section 4 discusses the experimental results from the application of our framework and several other clustering methods to four collections of text documents, related to four given queries, respectively. Finally, some concluding remarks are provided in Section 5.

2 RELATED WORK

Topic detection is traditionally considered as a clustering problem [3], due to the absence of training sets. The clustering task usually involves feature selection [4], spectral clustering [5] and k-means oriented [3] techniques, assuming mainly that the number of topics to be discovered is known a priori and there is no noise, i.e. news items that do not belong to any of the news clusters. Latent Dirichlet Allocation (LDA) is a popular approach for topic modelling for a given number of topics k [6]. LDA has been generalized to nonparametric Bayesian approaches, such as the hierarchical Dirichlet process [7] and DP-means [8], which predict the number of topics k. The extraction of the correct number of topics is equivalent to the estimation of the correct number of clusters in a dataset. The majority vote among 30 clustering indices has been proposed in [9] as an indicator for the number of clusters in a dataset. In contrast, we propose an alternative majority vote among 10 realizations of the “DBSCAN-Martingale”, which is a modification of the DBSCAN algorithm [10] with parameters the density level ε and a lower bound for the minimum number of points per cluster. However, the DBSCAN-Martingale [2] regards the density level ε as a random variable and the clusters are progressively extracted. We consider the general case, where the number of topics to be discovered is unknown and it is possible to have news articles which are not assigned to any topic.

Graph-based methods for event detection and multimodal clustering in social media streams have appeared in [11], where a graph clustering algorithm is applied on the graph of items. The decision, whether to link two items or not, is based on the output of a classifier, which assigns or not, the candidate items in the same

[1] Information Technologies Institute, CERTH, Thessaloniki, Greece, email: {heliasgj, dliparas, stefanos, ikom}@iti.gr
[2] **http://www.multisensorproject.eu/**

cluster. Contrary to this graph-based approach, we cluster news items in an unsupervised way.

Density-based clustering does not require as input the number of topics. OPTICS [12] is very useful for the visualization of the cluster structure and for the optimal selection of the density level ε. The OPTICS-ξ algorithm [12] requires an extra parameter ξ, which has to be manually set in order to find "dents" in the OPTICS reachability plot. The automatic extraction of clusters from the OPTICS reachability plot, as an extension of the OPTICS-ξ algorithm, has been presented in [13] and has been outperformed by HDBSCAN [14] in several datasets of any nature. In the context of news clustering, however, we shall examine whether some of these density-based algorithms perform well on the topic detection problem and by comparing them with our DBSCAN-Martingale, in terms of the number of estimated topics. All the aforementioned methods, which do not require the number of topics to be known a priori, are combined with LDA in order to examine whether the use of DBSCAN-Martingale (combined with LDA) provides the most efficient assignment of news articles to topics.

3 TOPIC DETECTION USING CONCEPTS AND NAMED ENTITIES

The MULTISENSOR framework for topic detection, which is presented in Figure 1, is approached as a news clustering problem, where the number of topics needs to be estimated. The overall framework is based on textual features, namely concepts and named entities. The number of topics k is estimated by DBSCAN-Martingale and the assignment of news articles to topics is done using Latent Dirichlet Allocation (LDA).

LDA has shown great performance in text clustering, given the number of topics. However, in realistic applications, the number of topics is unknown to the system. On the other hand, DBSCAN does not require as input the number of clusters, but its performance in text clustering is very weak, due to the fact that it assigns too much noise to the news article collection and this results in very limited performance [2]. Moreover, it is difficult to find a unique density level that can output all clusters. Thus, we keep only the number of clusters using density-based clustering and the assignment of documents to topics is done by the well-performing LDA.

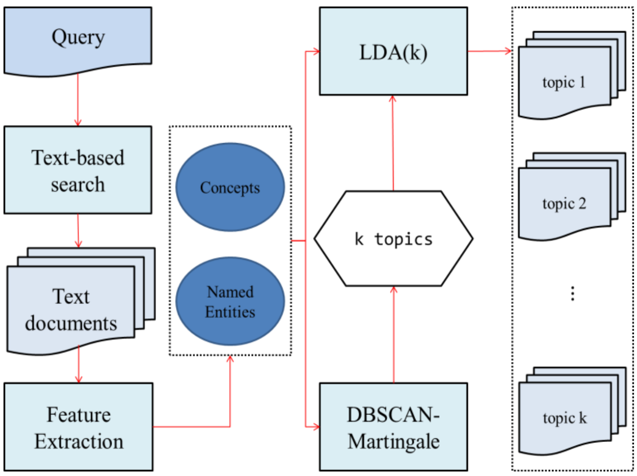


Figure 1. The MULTISENSOR topic detection framework using DBSCAN-Martingale and LDA

In our approach, the constructed DBSCAN-Martingale combines several density levels and is applied on high-level concepts and named entities. In the following, the construction of DBSCAN-Martingale is briefly reported.

3.1 The DBSCAN-Martingale

Given a collection of n news articles, density-based clustering algorithms output clustering vector C with values the cluster IDs $C[j]$ for each news item $j = 1,2,\dots,n$, where we denote by $C[j]$ the j-th element of a vector C. In case the j-th document is not assigned to any of the clusters, the j-th cluster ID is zero. Assuming that $C_{DBSCAN(\varepsilon)}$ is the clustering vector provided by the DBSCAN [10] algorithm for the density level ε, the problem is to combine the results for several values of ε, into one unique clustering result. To that end, a martingale construction has been presented in [2], where the density level ε is a random variable, uniformly sampled in a pre-defined interval.

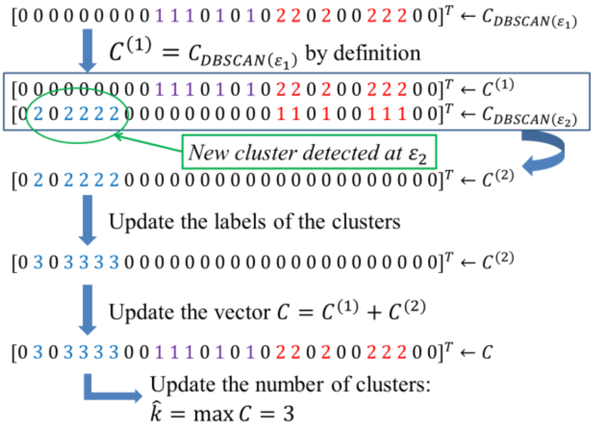

[0 0 0 0 0 0 0 0 0 1 1 1 0 1 0 1 0 2 2 0 2 0 0 2 2 2 0 0]^T ← C_DBSCAN(ε1)
C^(1) = C_DBSCAN(ε1) by definition
[0 0 0 0 0 0 0 0 0 1 1 1 0 1 0 1 0 2 2 0 2 0 0 2 2 2 0 0]^T ← C^(1)
[0 2 0 2 2 2 2 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 1 1 1 0 0]^T ← C_DBSCAN(ε2)
New cluster detected at ε2
[0 2 0 2 2 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]^T ← C^(2)
Update the labels of the clusters
[0 3 0 3 3 3 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]^T ← C^(2)
Update the vector C = C^(1) + C^(2)
[0 3 0 3 3 3 3 0 0 1 1 1 0 1 0 1 0 2 2 0 2 0 0 2 2 2 0 0]^T ← C
Update the number of clusters:
k̂ = max C = 3

Figure 2. One realization of the DBSCAN-Martingale with T = 2 iterations and 3 topics detected [2]

The DBSCAN-Martingale progressively updates the estimation of the number of clusters (topics), as shown in Figure 2, where 3 topics are detected in 2 iterations of the process. Due to the randomness in the selection of the density levels ε, it is likely that each realization of the DBSCAN-Martingale will output a random variable $\hat{k}$ as an estimation of the number of clusters. Hence, we allow 10 realizations $\widehat{k_1}, \widehat{k_2}, \dots, \widehat{k_{10}}$ and the final estimation of the number of clusters is the majority vote over them. An illustrative example of 5 clusters in the 2-dimensional plane is demonstrated in Figure 3.

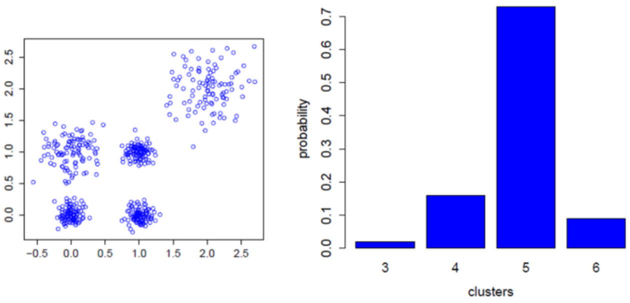


Figure 3. Example in the 2-dimensional plane and the histogram of results after 100 realizations of the DBSCAN-Martingale

In brief, the DBSCAN-Martingale is mathematically formulated as follows. Firstly, a sample of size T ε_t, $t = 1,2,\dots,T$ is randomly generated in $[0, \varepsilon_{max}]$, where ε_{max} is an upper bound for the density levels. The sample of ε_t, $t = 1,2,\dots,T$ is then sorted in increasing order. For each density level ε_t we find the corresponding clustering vectors $C_{DBSCAN(\varepsilon_t)}$ for all stages $t = 1,2,\dots,T$. In the first stage, all clusters detected by $C_{DBSCAN(\varepsilon_1)}$ are kept, corresponding to the lowest density level ε_1. In the second stage ($t = 2$), some of the detected clusters by $C_{DBSCAN(\varepsilon_2)}$ are new and some of them have also been detected by $C_{DBSCAN(\varepsilon_1)}$. In order to keep only the newly detected clusters, we keep only groups of numbers of the same cluster ID with size greater than $minPts$. Finally, the cluster IDs are relabelled and the maximum value of a clustering vector provides the number of clusters.

Complexity: The DBSCAN-Martingale requires T iterations of the DBSCAN algorithm, which runs in $\mathcal{O}(n \log n)$ if a tree-based spatial index can be used and in $\mathcal{O}(n^2)$ without tree-based spatial indexing [12]. Therefore, the DBSCAN-Martingale runs in $\mathcal{O}(Tn \log n)$ for tree-based indexed datasets and in $\mathcal{O}(Tn^2)$ without tree-based indexing. Our code[3] is written in R[4], using the dbscan[5] package, which runs DBSCAN in $O(n\ log\ n)$ with kd-tree data structures for fast nearest neighbor search.

3.2 Latent Dirichlet Allocation (LDA)

LDA assumes a Bag-of-Words (BoW) representation of the collection of documents and each topic is a distribution over terms in a fixed vocabulary. LDA assigns probabilities to words and assumes that documents exhibit multiple topics, in order to assign a probability distribution on the set of documents. Finally, LDA assumes that the order of words does not matter and, therefore, LDA is not applicable to word n-grams for $n \geq 2$, but can be applied to named entities and concepts. This input allows topic detection even in multilingual corpora, where n-grams are not available in a common language.

4 EXPERIMENTS

In this Section, we describe our dataset and evaluate our method.

4.1 Dataset description

A part of the present MULTISENSOR database (in which articles crawled from international news websites are stored) was used for the evaluation of our query-based topic detection framework. We use the retrieved results for a given query in order to cluster them into labelled clusters (topics) without knowing the number of clusters. The concepts and named entities are extracted using the DBpedia spotlight[6] online tool and the final concepts and named entities replaced the raw text of each news article. The final collection of text documents is available online[7].

The queries that were used for the experiments are the following:

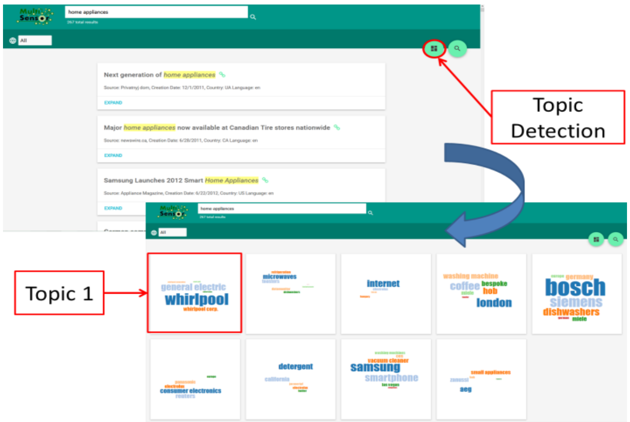
- energy crisis
- energy policy
- home appliances
- solar energy

It should be noted that the aforementioned queries are considered representative, with respect to the use cases addressed by the MULTISENSOR project. The output of our topic detection framework can be visualized in Figure 4 for the query "home appliances", where the retrieved results are clustered by 9 topics. The font size of the clusters' labels depends on the particular word probability within each cluster.

4.2 Evaluation results

In order to evaluate the clustering of the retrieved news articles, we use the average precision (AP), broadly used in the context of information retrieval, clustering and classification. A document d of a cluster C is considered relevant to C (true positive), if at least one concept associated with document d appears also in the label of cluster C. It should be noted that the labels of the clusters (topics) are provided by the concepts or named entities that have the highest probability (provided by LDA) within each topic. Precision is considered the fraction of relevant documents in a cluster and average precision is the average for all clusters of a query. Finally, we average the AP scores for all considered queries to obtain the Mean Average Precision (MAP).

We compared the clustering performance of the proposed topic detection framework, in which the DBSCAN-Martingale algorithm (for estimating the number of topics) and LDA (for assigning news articles to topics) are employed, against a variety of well-known clustering approaches, which were also combined with LDA for a fair comparison. DP-means is a Dirichlet process and we used its implementation in R[8]. HDBSCAN is a hierarchical DBSCAN approach, which uses the "excess-of-mass" (EOM) approach to find the optimal cut. Nbclust is a majority vote of the first 16 indices, which are all described in detail in [9].



home appliances

267 total results

Next generation of home appliances

Source: Privatnyj dom, Creation Date: 12/1/2011, Country: UA Language: en

EXPAND

Major home appliances now available at Canadian Tire stores nationwide

Source: newswire.ca, Creation Date: 6/28/2011, Country: CA Language: en

EXPAND

Samsung Launches 2012 Smart Home Appliances

Source: Appliance Magazine, Creation Date: 6/22/2012, Country: US Language: en

EXPAND

Figure 4. Demonstration of the MULTISENSOR topic detection framework

[3] https://github.com/MKLab-ITI/topic-detection

[4] https://www.r-project.org/

[5] https://cran.r-project.org/web/packages/dbscan/index.html

[6] https://dbpedia-spotlight.github.io/demo/

[7] http://mklab2.iti.gr/project/query-based-topic-detection-dataset

[8] https://github.com/johnmyleswhite/bayesian_nonparametrics

Table 1. Average Precision (± standard deviation) and Mean Average Precision over 10 runs of LDA using the estimated number of topics

Index + LDA	**energy crisis**	**energy policy**	**home appliances**	**solar energy**	**MAP**
CH	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
Duda	0.4498±0.0671	**0.5534±0.0457**	0.4299±0.0237	0.4484±0.0067	0.4703
Pseudo t^2	0.4498±0.0671	**0.5534±0.0457**	0.4299±0.0237	0.4484±0.0067	0.4703
C-index	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
Ptbiserial	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
DB	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
Frey	0.3541±0.0181	0.3911±0.0033	0.3745±0.064	0.4484±0.0067	0.3920
Hartigan	0.5938±0.0502	0.5336±0.0375	0.5942±0.0282	0.5961±0.0347	0.5794
Ratkowsky	0.5357±0.0151	0.5371±0.0357	0.4962±0.0721	0.5375±0.0446	0.5266
Ball	0.4207±0.0093	0.4501±0.0021	0.4975±0.016	0.4464±0.0614	0.4536
McClain	0.5786±0.0425	0.5371±0.0357	0.3745±0.064	0.5961±0.0347	0.5215
KL	0.5786±0.0425	0.5371±0.0357	0.5701±0.0145	0.5961±0.0347	0.5704
Silhouette	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
Dunn	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
SDindex	0.3541±0.0181	0.3911±0.0033	0.5942±0.0282	0.4484±0.0067	0.4469
SDbw	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
NbClust	0.5786±0.0425	0.5371±0.0357	0.5942±0.0282	0.5961±0.0347	0.5765
DP-means	0.3541±0.0181	0.3911±0.0033	0.3745±0.064	0.4484±0.0067	0.3920
HDBSCAN-EOM	0.4498±0.0671	0.3911±0.0033	0.5951±0.0184	0.5375±0.0446	0.4933
DBSCAN-Martingale	**0.7691±0.0328**	**0.5534±0.0457**	**0.6115±0.0225**	**0.6073±0.0303**	**0.6353**

Table 2. Estimation of the number of topics in the MULTISENSOR queries

Index	**energy crisis**	**energy policy**	**home appliances**	**solar energy**
CH	12	8	15	15
Duda	**4**	**4**	**3**	**2**
Pseudo t^2	4	4	3	2
C-index	12	8	15	15
Ptbiserial	12	8	15	15
DB	12	8	15	15
Frey	2	2	2	2
Hartigan	11	7	15	15
Ratkowsky	7	8	5	5
Ball	3	3	3	3
McClain	12	8	2	15
KL	12	8	11	15
Silhouette	12	8	15	15
Dunn	12	8	15	15
SDindex	2	2	15	2
SDbw	12	8	15	15
NbClust	12	8	15	15
DP-means	2	2	2	2
HDBSCAN-EOM	4	2	10	5
DBSCAN-Martingale	6	4	9	10

The AP scores per query and the MAP scores per method over 10 runs of LDA are displayed in Table 1, for each estimation of the number of topics combined with LDA. In addition, the numbers of news clusters estimated by the considered clustering indices for each query are presented in Table 2. Looking at Table 1, we observe a relative increase of 9.65% in MAP, when our topic detection framework is compared to the second highest MAP score (by Hartigan+LDA) and a relative increase of 10.20%, when compared to the most recent approach (NbClust+LDA).

In general, the proposed topic detection framework outperforms all the considered clustering approaches both in terms of AP (within each query) and in terms of MAP (overall performance for all queries), with the exception of the "energy policy" query, where the performance of our framework is matched by that of the Duda and Pseudo t^2 clustering indices.

Finally, we evaluated the time performance of the DBSCAN-Martingale method and we selected several baseline approaches in order to compare their processing time with that of our approach. In Figure 5, the number of news clusters is estimated for $T = 5$ iterations for the DBSCAN-Martingale and for maximum number of clusters set to 15 for the indices Duda, Pseudo t^2, Silhouette, Dunn and SDindex. We observe that DBSCAN-Martingale is faster than all other methods. Even when it is applied to 500 documents, it is able to reach a decision about the number of clusters in approximately 0.4 seconds.

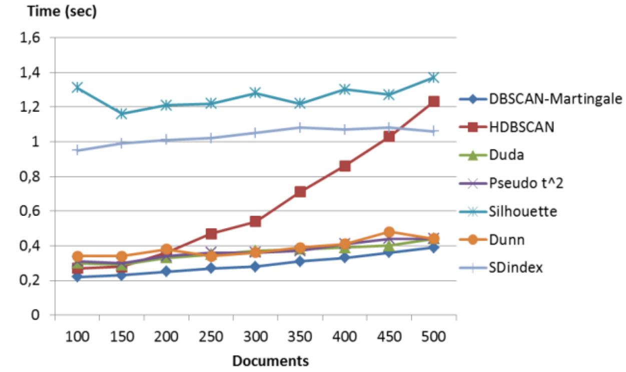


Figure 5. Time performance of DBSCAN-Martingale and several baseline approaches to estimate the number of news clusters

5 CONCLUSIONS

In this paper, we have presented a hybrid topic detection framework, developed for the purposes of the MULTISENSOR project. Given a query-based search, the framework clusters the retrieved results by topic, without the need to know the number of topics a priori. The framework employs the recently introduced DBSCAN-Martingale method for efficiently estimating the number of news clusters, coupled with Latent Dirichlet Allocation for assigning the news articles to topics. Our topic detection framework relies on high-level textual features that are extracted from the news articles, namely textual concepts and named entities. In addition, it is multimodal, since it fuses more than one sources of information from the same multimedia object. The query-based topic detection experiments have shown that our framework outperforms several well-known clustering methods, both in terms of Average Precision and Mean Average Precision. A direct comparison, by means of time performance, has shown that our approach is faster than several well-performing methods in the estimation of the number of clusters, given as input the same number of query-based retrieved news articles.

As future work, we plan to investigate the behavior of our framework by introducing additional modalities/features, examine the application of alternative (other than LDA) text clustering approaches, as well as investigate the extraction of language-agnostic concepts and named entities, something that could provide multilingual capabilities to our topic detection framework.

ACKNOWLEDGEMENTS

This work was supported by the projects MULTISENSOR (FP7-610411) and KRISTINA (H2020-645012), funded by the European Commission.

REFERENCES

[1] J. Allan (Ed.), 'Topic detection and tracking: event-based information organization', vol. 12, Springer Science & Business Media, (2012).

[2] I. Gialampoukidis, S. Vrochidis and I. Kompatsiaris, 'A hybrid framework for news clustering based on the DBSCAN-Martingale and LDA', In: Perner, P. (Ed.) Machine Learning and Data Mining in Pattern Recognition, LNAI 9729, pp. 170-184, (2016).

[3] C. C. Aggarwal and C. Zhai, 'A survey of text clustering algorithms', In Mining Text Data, pp. 77-128, Springer US, (2012).

[4] M. Qian and C. Zhai, 'Unsupervised feature selection for multi-view clustering on text-image web news data', In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1963-1966, ACM, (2014).

[5] A. Kumar and H. Daumé, 'A co-training approach for multi-view spectral clustering', In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 393-400, (2011).

[6] D. M. Blei, A. Y. Ng and M. I. Jordan, 'Latent dirichlet allocation', the Journal of machine Learning research, vol. 3, pp. 993-1022, (2003).

[7] Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei, 'Hierarchical dirichlet processes', Journal of the american statistical association, 101(476), (2006).

[8] B. Kulis and M. I. Jordan, 'Revisiting k-means: New algorithms via Bayesian nonparametrics', arXiv preprint arXiv:1111.0352, (2012).

[9] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs, 'NbClust: an R package for determining the relevant number of clusters in a data set', Journal of Statistical Software, 61(6), pp. 1-36, (2014).

[10] M. Ester, H. P. Kriegel, J. Sander and X. Xu, 'A density-based algorithm for discovering clusters in large spatial databases with noise', In Kdd, 96(34), pp. 226-231, (1996).

[11] G. Petkos, M. Schinas, S. Papadopoulos and Y. Kompatsiaris, 'Graph-based multimodal clustering for social multimedia', Multimedia Tools and Applications, 1-23, (2016).

[12] M. Ankerst, M. M. Breunig, H. P. Kriegel and J. Sander, 'OPTICS: ordering points to identify the clustering structure', In ACM Sigmod Record, 28(2), pp. 49-60, ACM, (1999).

[13] J. Sander, X. Qin, Z. Lu, N. Niu and A. Kovarsky, 'Automatic extraction of clusters from hierarchical clustering representations', In Advances in knowledge discovery and data mining, pp. 75-87, Springer Berlin Heidelberg, (2003).

[14] R. J. Campello, D. Moulavi and J. Sander, 'Density-based clustering based on hierarchical density estimates', In Advances in Knowledge Discovery and Data Mining, pp. 160-172, Springer Berlin Heidelberg, (2013).