Combining Latent Dirichlet Allocation and K-Means for Documents Clustering: Effect of Probabilistic Based Distance Measures

Quang Vu Bui[1,2], Karim Sayadi[2(✉)], Soufian Ben Amor[3,4], and Marc Bui[2]

[1] Hue University of Sciences, Hue, Vietnam
[2] CHArt Laboratory EA 4004, EPHE, PSL Research University, Paris, France
karim.sayadi@ephe.sorbonne.fr
[3] LI-PARAD Laboratory, University of Versailles-Saint- Quentin-en-Yvelines, Versailles, France
[4] Paris-Saclay University, Paris, France

Abstract. This paper evaluates through an empirical study eight different distance measures used on the LDA + K-means model. We performed our analysis on two miscellaneous datasets that are commonly used. Our experimental results indicate that the probabilistic-based distance measures are better than the vector based distance measures including Euclidean when it comes to cluster a set of documents in the topic space. Moreover, we investigate the implication of the number of topics and show that K-means combined to the results of the Latent Dirichlet Allocation model allows us to have better results than the LDA + Naive and Vector Space Model.

Keywords: Latent Dirichlet Allocation · Topic modeling · Document clustering · K-means · Similarity measure · Probabilistic-based distance · Clustering evaluation

1 Introduction

Clustering a set of documents is a standard problem addressed in data mining, machine learning, and statistical natural language processing. Document clustering can automatically organize many documents into a small number of meaningful clusters and find latent structure in unlabeled document collections.

K-means is one of the most used partitioned-based clustering algorithms. It became popular among information retrieval tasks [12]. For clustering a set of documents with K-means, each document is firstly quantified as a vector where each component indicates a corresponding feature in the document. Then, a distance is used to measure the difference between two documents. The collection of documents is represented by a sparse and high-dimensional matrix. The use of this matrix raises an issue known as the "curse of dimensionality" [14]. Thus, using K-means require reducing the documents dimensionality and using a "good" distance measure to get the most accurate clusters.

N.T. Nguyen et al. (Eds.): ACIIDS 2017, Part I, LNAI 10191, pp. 248–257, 2017.
DOI: 10.1007/978-3-319-54472-4_24

In our work, we first reduce the dimensionality by decomposing the document matrix into latent components using the Latent Dirichlet Allocation (LDA) [2] method. Each document is represented by a probability distribution of topics and each topic is characterized by a probability distribution over a finite vocabulary of words. We use the probability distribution of topics as the input for K-means clustering. This approach called LDA + K-means was proposed by [3,17]. We note that [17] proposed LDA + K-means but only used Euclidean distance.

We then compare the efficiency of eight distance measures [5]. These measures are based on two approaches: (i) Vector based approach (VBM) with Euclidean distance, Sørensen distance, Tanimoto distance, Cosine distance and (ii) Probabilistic-based approach (PBM) with Bhattacharyya distance, Probabilistic Symmetric χ^2 divergence, Jensen-Shannon divergence, Taneja divergence.

In order to come up with a sound conclusion, we have performed an empirical evaluation of the eight distance measures according to a labeled clustering. We compared the clusters with the two evaluation criteria: Adjusted Rand Index (ARI) [9] and Adjusted Mutual Information (AMI) [16]. We used two common datasets in the NLP community: the 20NewsGroup dataset contains newsgroup posts and the WebKB contains texts extracted from web pages.

Our experiments can be compared to the work of [8,11,17]. The key differences are the following: In comparison with the VBM we conducted our experiments with a PBM, we show that in the case of LDA + K-means where the input is a probability distribution the use of PBM leads to better results. Then, our results show that the Euclidean distance may not be suitable for this kind of application. Finally, by evaluating the results of the VBM and PBM with ARI and AMI criteria we have investigated the implication of the number of topics in the clustering processing.

This paper is organized as follows. The next section describes the methodology in which we present K-means algorithms and document clustering, similarity measures in probabilistic spaces and evaluation indexes used in the experiments. We explain the experiment, discuss the results in Sect. 3 and also conclude our work in Sect. 4.

2 Methodology

2.1 Document Clustering

Vector Space Model. Most current document clustering methods choose to view text as a bag of words. In this method, each document is represented by word-frequency vector $d_{wf} = (wf_1, wf_2, \ldots, wf_n)$, where wf_i is the frequency of the ith word in the document. This gives the model its name, the vector space model (VSM) [15].

The two disadvances of VSM are the high dimensionality because of the high number of unique terms in text corpora and insufficient to capture all semantics. Latent Dirichlet Allocation [2] proposed a good solution to solve these issues.

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) [2] is a generative probabilistic model for topic discovery. In LDA, each document may be considered as a mixture of different topics and each topic is characterized by a probability distribution over a finite vocabulary of words. The generative model of LDA, described with the probabilistic graphical model in Fig. 1, proceeds as follows:

1. Choose distribution over topics θ_i from a Dirichlet distribution with parameter α for each document.
2. Choose distribution over words ϕ_k from a Dirichlet distribution with parameter β for each topic.
3. For each of the word positions i, j:
   3.1. Choose a topic $z_{i,j}$ from a Multinomial distribution with parameter θ_i
   3.2. Choose a word $w_{i,j}$ from a Multinomial distribution with parameter $\phi_{z_{i,j}}$

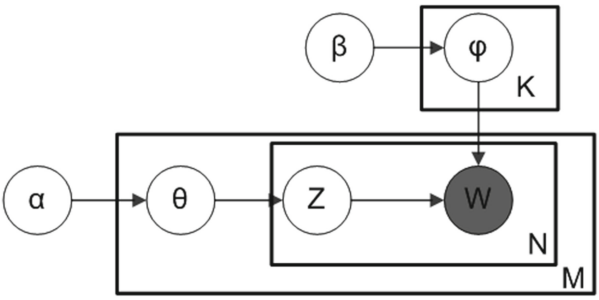


Fig. 1. Probabilistic graphical model of LDA

For posterior inference, we need to solve the following equation:

$$p(\theta, \phi, z|w, \alpha, \beta) = \frac{p(\theta, \phi, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}$$

There are some inference algorithms available including variational inference used in the original paper [2] and Gibbs Sampling. Please refer to the work of [1] for more details.

K-Means Algorithm. K-means which proposed by Forgy [6] is one of the most popular clustering algorithms. It provides a simple and easy way to classify objects in k groups fixed a priori. The basic idea is to define k centroids and then assign objects to the nearest centroid. A loop has been generated. In each step, we need to re-calculate k new centroids and re-assign objects until no more changes are done. The algorithm works as follows:

1. Selecting k initial objects called centroids of the k clusters.
2. Assigning each object to the cluster that has the closest centroid.
3. Computing the new centroid of each cluster.
4. Repeat step 2 and 3 until the objects in any cluster do no longer change.

2.2 Combining LDA and K-Means

The output of LDA is two probability distributions: the document-topic distribution θ and the word-topic distribution ϕ. To use as much as possible information from LDA result, we can combine Latent Dirichlet Allocation and K-means, denoted LDA + K-means, by using document-topic distributions θ extracted from LDA as the input for K-means clustering algorithms. For a matter of space, we invite the readers to find more details in the work of [3].

2.3 Similarity Measures

Since LDA represents documents as probability distributions, we need to consider the "good" way to choose a distance or similarity measure for comparing two probability distributions. Eight distances families as categorized by [5] were used in K-means + LDA. These families can be divided into two groups:

- Vector-Based Measurements (VBM): Euclidean distance, Sørensen distance, Tanimoto distance, Cosine distance
- Probabilistic-Based Measurements (PBM): Bhattacharyya distance, Probabilistic Symmetric χ^2 divergence, Jensen-Shannon divergence, Taneja divergence

Let $A = (a_1, a_2, \ldots, a_k)$ and $B = (b_1, b_2, \ldots, b_k)$ be two vectors with k dimensions. The eight distances between A and B are defined as:

Euclidean distance: $d_{Euc} = \sqrt{\sum_{i=1}^{k} |a_i - b_i|^2}$

Sørensen distance: $d_{Sor} = \frac{\sum_{i=1}^{k} |a_i - b_i|}{\sum_{i=1}^{k} (a_i + b_i)}$

Tanimoto distance: $d_{Tani} = \frac{\sum_{i=1}^{k} (max(a_i, b_i) - min(a_i, b_i))}{\sum_{i=1}^{k} max(a_i, b_i)}$

Cosine distance: $d_{Cos} = 1 - Sim_{Cos} = 1 - \frac{\sum_{i=1}^{k} a_i b_i}{\sqrt{\sum_{i=1}^{k} {a_i}^2} \sqrt{\sum_{i=1}^{k} {b_i}^2}}$

Jensen-Shannon Divergence. The Jensen-Shannon (JS) divergence, known as a total divergence to the average, is based on Kullback-Leibler (KL) divergence, which is related to Shannon's concept of uncertainty or "entropy" $H(A) = \sum_{i=1}^{k} a_i ln a_i$.

$$d_{JS} = \frac{1}{2} \sum_{i=1}^{k} a_i ln\left(\frac{2a_i}{a_i + b_i}\right) + \frac{1}{2} \sum_{i=1}^{k} b_i ln\left(\frac{2b_i}{a_i + b_i}\right)$$

Bhattacharyya Distance. Bhattacharyya distance is a divergence-type measure between distributions, defined as,

$$d_{Bhat} = -ln \sum_{i=1}^{k} \sqrt{a_i b_i}$$

Probabilistic Symmetric χ^2 Divergence. Probabilistic Symmetric χ^2 divergence is a special case of χ^2 divergence. It is a combination of Pearson χ^2 divergence and Newman χ^2 divergence.

$$d_{PChi} = 2 \sum_{i=1}^{k} \frac{(a_i - b_i)^2}{a_i + b_i}$$

Taneja Divergence. Taneja divergence is a combination between KL divergence and Bhattacharyya distance, using KL-divergence with $a_i = \frac{a_i + b_i}{2}, b_i = \sqrt{a_i b_i}$

$$d_{TJ} = \sum_{i=1}^{k} (\frac{a_i + b_i}{2}) ln (\frac{a_i + b_i}{2\sqrt{a_i b_i}})$$

2.4 Evaluation Methods

For each dataset, we obtained a clustering result from the K-means algorithm. To measure the quality of the clustering results, we used two evaluation indexes: Adjusted Rand Index (ARI) [9] and Adjusted Mutual Information (AMI) [16], which are widely used to evaluate the performance of unsupervised learning algorithms.

Adjusted Rand Index: Adjusted Rand Index (ARI) [9], an adjusted form of Rand Index (RI), is defined as:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\circ}}{2} \sum_j \binom{n_{\circ j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\circ}}{2} + \sum_j \binom{n_{\circ j}}{2}\right] - \left[\sum_i \binom{n_{i\circ}}{2} \sum_j \binom{n_{\circ j}}{2}\right] / \binom{n}{2}} \tag{1}$$

where $n_{ij}, n_{i\circ}, n_{\circ j}, n$ are values from the contingency Table 1.

Adjusted Mutual Information. The Adjusted Mutual Information (AMI) [16], an adjusted form of mutual information (MI), is defined:

$$AMI(P,Q) = \frac{MI(P,Q) - E\{MI(P,Q)\}}{\max\{H(P), H(Q)\} - E\{MI(P,Q)\}} \tag{2}$$

where

$$H(P) = -\sum_{i=1}^{k} \frac{n_{i\circ}}{n} \log \frac{n_{i\circ}}{n}; MI(P,Q) = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}}{n} \log \frac{n_{ij}/n}{n_{i\circ} n_{\circ j}/n^2}.$$

Table 1. The Contingency Table, $n_{ij} = |P_i \cap Q_j|$

$P \setminus Q$	Q_1	Q_2	$\cdots$	Q_l	Sums
P_1	n_{11}	n_{12}	$\cdots$	n_{1l}	$n_{1\circ}$
P_2	n_{21}	n_{22}	$\cdots$	n_{2l}	$n_{2\circ}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
P_k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	$n_{k\circ}$
Sums	$n_{\circ 1}$	$n_{\circ 2}$	$\cdots$	$n_{\circ l}$	$\sum_{ij} n_{ij} = n$

Both ARI and AMI have a boundary above by 1. Higher values of ARI or AMI indicate more agreement between the two partitions. Please refer to the work of [9], [16] for more details.

3 Experiments and Results

3.1 Datasets

The proposed methodology is evaluated on 2 miscellaneous datasets that are commonly used for the NLP community regarding the task of document clustering. Table 2 describes some statistics about the used datasets. The 20Newsgroup collect has 18821 documents distributed across 20 different news categories. Each document corresponds to one article with a header that contains the title, the subject, and quoted text. The WebKB dataset contains 8230 web pages from the computer science department of different universities (e.g. Texas, Wisconsin, Cornell, etc.).

Table 2. Statistics of the datasets. Where #Docs refers to the number of documents in the dataset, #Classes refers to the number of classes in the dataset and < Class, > Class, refers to the minimum number of documents and the maximum number of document in a class.

Dataset	#Docs	#Classes	< Class	> Class
News20	18821	20	628	999
WebKB	8230	4	504	1641

3.2 Setup

In our experiments, we compared eight distances used with LDA + K-means divided into the two categories: the Probabilistic-Based Measurements (PBM) and the Vector-Based Measurements (VBM). We run LDA with Gibbs sampling method using the `topicmodels` R package[1]. The prior parameters α and β are

[1] https://cran.r-project.org/web/packages/topicmodels/index.html.

respectively set to 0.1 and 0.01. These parameters were chosen according to the state-of-the-art standards [7]. The number of iterations of the Gibbs sampling is set to 5000. The input number of topics for the 20NewsGroups dataset is set to 30 and for the WebKB dataset is set to 8. This number of topics will be confirmed in our experiments by testing different values. For each of the eight distances, we run the K-means 20 times with a maximum number of iterations equal to 1000. We compute the ARI and AMI on the results of each K-means iteration and report the average values.

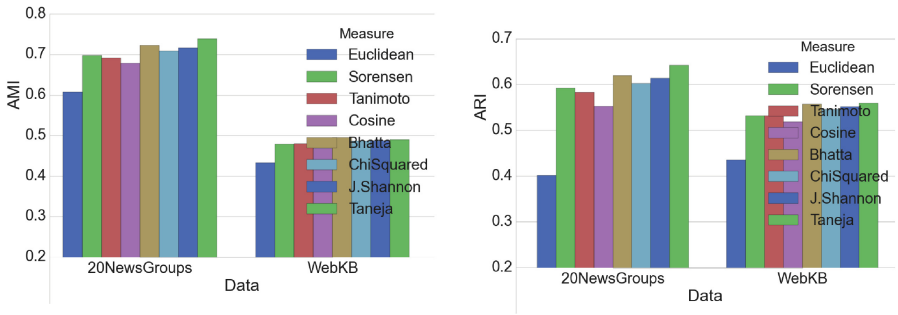
3.3 Results

Comparing Effectiveness of Eight Distance Measures for LDA + K-Means. The average values of the ARI and AMI are reported in Table 3. The average ARI and AMI values of the PBM group are better than the average values of the VBM group. We notice that the Euclidean distance has the worst results regarding the ARI and AMI criteria. In the PBM group, the best average values are obtained by the two distances Bhattacharyya and Taneja. Thus, we propose to work with Taneja or Bhattacharyya distance for LDA + K-means. For a better understanding of the results, we additionally provide a bar plot illustrated in Fig. 2.

Table 3. The average values of ARI, AMI for VSM, LDA Naive, LDA + K-means with eight different distance measures for two datasets

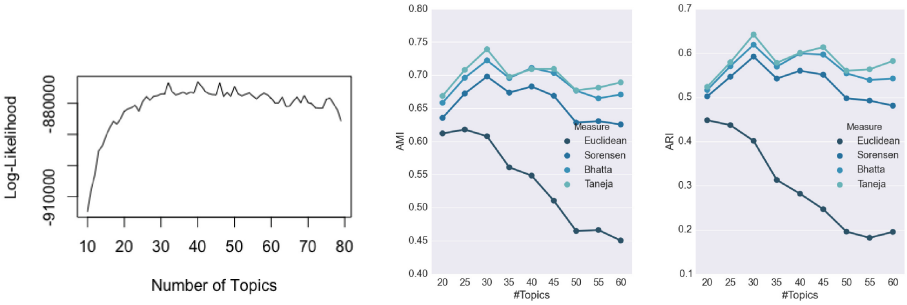
Distances	20NewsGroups		WebKB	
	ARI	AMI	ARI	AMI
Euclidean	0,402	0,608	0,436	0,432
Sorensen	0,592	0,698	0,531	0,479
Tanimoto	0,582	0,691	0,531	0,48
Cosine	0,552	0,678	0,519	0,468
Bhattacharyya	0,619	0,722	0,557	0,495
ChiSquared	0,602	0,708	0,545	0,487
JensenShannon	0,614	0,717	0,551	0,488
Taneja	0,642	0,739	0,559	0,489
VSM	0,128	0,372	0,268	0,335
LDA + Naive	0,434	0,590	0,171	0,197

The Role Played by the Number of Topics for LDA + K-Means. We chose the number of topics based on the Harmonic mean of Log-Likelihood (HLK) [4]. We notice in the Fig. 3(a), that the best number of topics are in the range of $[30, 50]$ of a maximum value of HLK. We run the LDA + K-means with a different number of topics and four distances: two from the PBM group, two from the VBM group including the Euclidean distance. We plot the evaluation with AMI and ARI in the Fig. 3(b) and (c).



(a) AMI (b) ARI

Fig. 2. The average values of ARI, AMI for LDA + K-means with eight different distance measures for two datasets

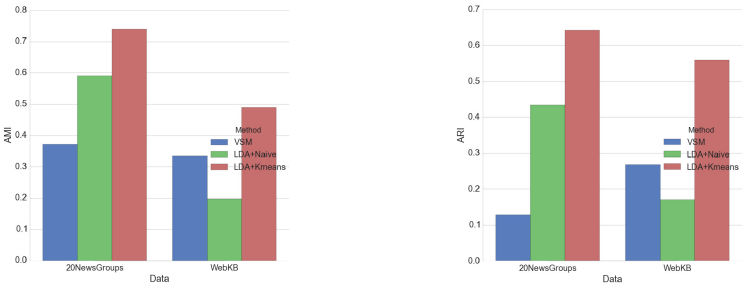


(a) Log-Likelihood (b) AMI (c) ARI

Fig. 3. The harmonic mean of the log-likelihood and ARI, AMI values with four distances for 20NG dataset with different # of topics.

As the number of topics increases, the LDA + K-means with Euclidean distance decreases in performance. The Euclidean distance is clearly not suitable for the LDA + K-means. The other three used distances (i.e. Sorensen, Bhattacharyya, and Taneja) kept a steady behavior with a slight advantage for the Taneja distance. This is due to the fact that these distance were defined for probability distribution and thus are more suitable for the kind of input provided by LDA. We notice that after 50 topics the performance of the three distances decreases.

Comparing LDA + K-Means, LDA + Naive, VSM. In order to study the role played by topic modeling, we compare three document clustering methods. The first is Vector space model (VSM) that uses a word-frequency vector $d_{wf} = (wf_1, wf_2, \ldots, wf_n)$, where wf_i is the frequency of the ith word in the document as input for K-means [13]. The second is proposed in [10], which



(a) AMI (b) ARI

Fig. 4. ARI, AMI values for three methods: VSM, LDA + Naive, LDA + K-means with Taneja distance computed on 20NGNewsGroups and WebKB datasets

considers each topic as a cluster. In fact, document-topic distribution θ can be viewed as a mixture proportion vector over clusters and thus can be used for clustering as follows. Suppose that x is a cluster, a document is assigned to x if $x = argmax_j \theta_j$. Note that this approach is a simple solution, usually referred to as a naive solution to combine topic modeling and document clustering. This approach is denoted in our experiments as LDA + Naive. The third one is the LDA + Kmeans with the probabilistic-based distance measure (eg. Bhattacharyya, Taneja). The results are plotted in Fig. 4, we notice that the LDA + Kmeans used with Taneja distance obtains the best average results for both of the used datasets.

4 Conclusion

In this paper, we compared the effect of eight distance or similarity measures represented to eight distance measure families for clustering document using LDA + K-means. Experiments on two datasets with two evaluation criteria demonstrate the fact that the efficiency of Probabilistic-based measurement clustering is better than the Vector based measurement clustering including Euclidean distance. Comparing among LDA + K-means, LDA + Naive, Vector Space Model, the experiments also show that if we choose the suitable value of a number of topic for LDA and Probabilistic-based measurements for K-means, LDA + K-means can improve the effect of clustering results.

References

1. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Bui, Q.V., Sayadi, K., Bui, M.: A multi-criteria document clustering method based on topic modeling and pseudoclosure function. Informatica **40**(2), 169–180 (2016)

4. Buntine, W.: Estimating likelihoods for topic models. In: Zhou, Z.-H., Washio, T. (eds.) ACML 2009. LNCS (LNAI), vol. 5828, pp. 51–64. Springer, Heidelberg (2009). doi:10.1007/978-3-642-05224-8_6
5. Cha, S.-H.: Comprehensive survey on distance/similarity measures between probability density functions. City **1**(2), 1 (2007)
6. Gordon, A.: Classification. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2nd edn. CRC Press, Boca Raton (1999)
7. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. **101**(Suppl 1), 5228–5235 (2004)
8. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC 2008), Christchurch, New Zealand, pp. 49–56 (2008)
9. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
10. Lu, Y., Mei, Q., Zhai, C.: Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Inf. Retrieval **14**(2), 178–203 (2010)
11. Maher, K., Joshi, M.S.: Effectiveness of different similarity measures for text classification and clustering. Int. J. Comput. Sci. Inf. Technol. **7**(4), 1715–1720 (2016)
12. Manning, C.D., Raghavan, P.: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009)
13. Modha, D.S., Spangler, W.S.: Feature weighting in k-means clustering. Mach. Learn. **52**(3), 217–237 (2003)
14. Pestov, V.: On the geometry of similarity search: dimensionality curse and concentration of measure. Inf. Process. Lett. **73**(1), 47–51 (2000)
15. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**(5), 513–523 (1988)
16. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
17. Xie, P., Xing, E.P.: Integrating Document Clustering and Topic Modeling, September 2013. arXiv:1309.6874