The Clustering-Based Initialization for Non-negative Matrix Factorization in the Feature Transformation of the High-Dimensional Text Categorization System: A Viewpoint of Term Vectors

Le Nguyen Hoai Nam(✉) and Ho Bao Quoc

VNUHCM - The University of Science, Ho Chi Minh City, Vietnam
{lnhnam,hbquoc}@fit.hcmus.edu.vn

Abstract. Due to the non-negativity of the matrix factors, Non-negative Matrix Factorization (NMF) is favorable for transforming a high-dimensional original Terms-Documents matrix into a lower-dimensional semantic Concepts-Documents matrix in the text categorization. With the iterative nature of all NMF algorithms, the NMF matrix factors need initializing. In this paper, we propose a clustering-based method for initializing the NMF according to the term vectors instead of the document vectors as the previous researches.

Keywords: Feature transformation · Matrix factorization · Text categorization

1 Introduction

Document categorization is a process that automatically classifies a given collection of documents into the predefined categories [21]. In this task, each document is converted to a Bag-of-Words (BoW) vector, so a corpus can be presented by a Terms-Documents matrix [19]. However, in a large corpus, the number of terms highly increases. This leads to the risk of irrelevant features, sparse vectors, and over-fitting producing the negative effects on the categorization [21]. In addition, the phenomenon of synonymy and polysemy has become more common when the corpus is larger.

Feature Transformation (FT) [14] is one of the main techniques to solve the problems in the high-dimensional text categorization. In an FT, all of the original terms join together to build new features, also known as semantic concepts. It is mainly concerned with a Low-Rank Approximation (LRA) to replace a high-dimensional original Terms-Documents matrix with a lower-dimensional semantic Concepts-Documents matrix. The document presentation under a new axis system being the semantic concepts helps deal with the issue of synonymy and polysemy [6]. A commonly used LRA is Singular Value Decomposition (SVD) [8]. However, the SVD cannot guarantee the non-negativity of the output matrices, although the Terms-Documents input matrix is inherently non-negative. After performing the SVD, the negative elements make the corpus presentation hard to be interpreted [9, 23]. Non-negative Matrix Factorization (NMF) [12] addresses this shortcoming by a non-negativity constraint on the matrix

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 511–522, 2017.
DOI: 10.1007/978-3-319-67008-9_40

approximation. Due to the iterative nature of all NMF algorithms, the NMF matrix factors need initializing. The previous researches [4, 24, 26] interpret the NMF matrix factors under the viewpoint of the column (document) vectors. Then, they are initialized according to this viewpoint by clustering the document vectors, i.e. the column vectors of the Terms-Documents matrix. However, an arising question is how the clustering on the document vectors initializes itself. In this case, it is very difficult to efficiently and quickly select which documents in a large corpus to be the first centroids for clustering.

To avoid this difficulty, we change viewing the NMF matrix factors from the column (document) vectors to the row (term) vectors. Under the row (term) viewpoint, we analyze the interpretation of the NMF matrix factors which is very different from that under the column (document) viewpoint. Therefore, we customize the idea of a clustering-based method to initialize the NMF matrix factors according to the row (term) vectors. The effectiveness of our method motivates further studies on the text FT based on the NMF not only at the NMF initialization stage but also the other NMF stages by interpreting the NMF matrix factors looking at their row (term) vectors.

2 Related Work

2.1 Non-negative Matrix Factorization (NMF)

NMF [12] is used to approximate a non-negative Terms-Documents matrix, called $X{:}R^{+}_{m\,x\,n}$, to the product of two non-negative matrices, called $W : R^{+}_{m\,x\,r}$ and $H : R^{+}_{r\,x\,n}$, with rank $r << min\{m, n\}$. For NMF, every original document, i.e. a column vector of the Terms-Documents matrix $X(X_{col_i} i = 1 \ldots n)$, can be reconstructed as follows:

$$X \approx W.H \Leftrightarrow X_{col_i} \approx W.H_{col_i} \Leftrightarrow X_{col_i} \approx \sum\nolimits_{j=1}^{r} (H_{ji}.W_{col_j}) \tag{1}$$

Specifically, the i^{th} original document vector (X_{col_i}) can be reconstructed by a linear combination of all the column vectors of W $(W_{col_j} j = 1 \ldots r)$ with the coefficients contained in the i^{th} column vector of H (H_{col_i}). Due to the non-negativity of W and H, the column vectors of W can be considered as the document basis vectors, and every column vector of H shows the real (only additive) coordinates of the corresponding original document vector with respect to the new axes being these document basis vectors. Therefore, under the viewpoint of the column vectors, W and H are addressed as a *document basis matrix* and a *document coordinate matrix*. This characteristic of the NMF is called the "parts-based presentation" because it shows an additive combination of the non-negative parts for constructing data [12, 23]. It makes sense to the analysis on real-world data in general and textual data in particular [23].

2.2 Text Feature Transformation Based on the NMF

A Feature Transformation (FT) [14] aims at not only reducing the feature space dimension but also creating more meaningful features by combining all the original features. It is to project the original document vectors onto a low-dimensional semantic subspace.

With W $(m\,x\,r)$ and H $(r\,x\,n)$ obtained from the NMF on the Terms-Documents matrix X $m\,x\,n)$, if the semantic subspace is spanned by the basis vectors being the column vectors of W, the i^{th} original document $(X_{col_i} : m\,x\,1)$ is now presented by the i^{th} column vector of H $(H_{col_i} : r\,x\,1)$, i.e. the projection of X_{col_i} onto the semantic subspace. In other words, an FT based on the NMF transforms the high-dimensional original Terms-Documents matrix X $(m\ x\ n)$ into the lower-dimensional semantic Concepts-Documents matrix H $(r\ x\ n)$ under the semantic basis W $(m\ x\ r)$.

Another typical FT is Latent Semantic Indexing (LSI) [7]. The LSI is based on a Singular Value Decomposition (SVD) [8] on the corpus matrix and then considers eigenvectors as the basis vectors of the semantic space for the transformation. However, the SVD cannot guarantee the non-negativity of its eigenvectors. Therefore, compared with the NMF, the SVD factors are less meaningful in text domain [23].

2.3 NMF Initialization

In NMF algorithms, W and H are iteratively updated to decrease an approximation error of X to $(W.H)$ [5, 12]. The most natural way for constructing this error function is to use Euclidean distance. Table 1 shows the general structure of a NMF algorithm.

Table 1. The general structure of a NMF algorithm.

Input: $X: R^{+}_{m\,x\,n}$; NMF rank $r << min\{m,n\}$. **Output**: $W: R^{+}_{m\,x\,r}$ and $H: R^{+}_{r\,x\,n}$.
Initialize W and H ($W^{(0)}$ and $H^{(0)}$).
While (Not satisfying **the convergence criterion**): **Update** W and H.

Due to the iterative characteristic, the NMF approximation error has a tendency to converge on a local minimum instead of a global minimum as expected [12]. It directly depends on the initial values of W and H, called $W^{(0)}$ and $H^{(0)}$. A good initialization leads the NMF to a faster convergence and better error at convergence [2, 4]. As presented above, all the column vectors of W $(W_{col_j}\ j = 1...r)$ play the role of the document basis vectors. They can easily be associated with the cluster centroids obtained from clustering the original document vectors. Therefore, a *clustering-based NMF initialization* [4, 24, 26] clusters the document original vectors $(X_{col_i}\ i = 1...n)$, and then utilizes the cluster centroids as the column vectors of $W^{(0)}$ $(W^{(0)}_{col_j}\ j = 1...r)$. Based on Eq. (1), the initial document coordinate matrix $H^{(0)}$ showing the relation between each original document $(X_{col_i} i = 1...n)$ and each cluster centroid (initial document basis) $W^{(0)}$ $(W^{(0)}_{col_j} j = 1...r)$. Specifically, $H^{(0)}_{ji}$ is 1 or 0 indicating whether X_{col_i} belongs to the cluster $W^{(0)}_{col_j}$ or not. Furthermore, an NMF can use the output factors of other matrix factorizations for its initialization. NNDSVD [2] implements a *factorization-based NMF initialization* by two SVDs.

3 Motivation

For the clustering-based NMF initialization, the commonly used clustering algorithms are the K-means (KM) and Fuzzy C-Means (FCM). However, one of the biggest challenges these clustering algorithms face is to determine a starting centroid for each cluster. In the KM-Clustering-based NMF initialization [24] and FCM-Clustering-based NMF initialization [26], the used clustering algorithm is started with cluster centroids selected at random among the original documents. However, the randomization makes the NMF non-deterministic. Recently, [4] implements a clustering-based NMF initialization with the Subtractive Clustering (SC), i.e. a clustering algorithm of no cluster centroid initialization. For SC, every obtained cluster centroid is just one of the original documents. If an original document is used as an initial document basis, the distance between the initial document basis and the true document basis is too far.

To improve a clustering-based NMF initialization, the clustering algorithm should begin to run with the initial cluster centroids being important documents instead of random documents. In this case, it is necessary to select important documents from the corpus simply and fast. However, this task is neither highly efficient nor inexpensive, especially with a large corpus. We realize that if X, W, and H are only interpreted under the viewpoint of the column (document) vectors, and W and H are then initialized according to this viewpoint, it is too difficult to define better starting centroids of a clustering algorithm when it used to initialize the NMF. Therefore, we introduce a new interpretation of the NMF by looking at the row (term) vectors as follows:

$$X \approx W.H \Leftrightarrow X_{row_i} \approx W_{row_i}.H \Leftrightarrow X_{row_i} \approx \sum_{j=1}^{r} (W_{ij}.H_{row_j}) \tag{2}$$

Concretely, the i^{th} original term vector, i.e. the i^{th} row vector of X (X_{row_i}), is constructed by a linear combination of the row vectors of H $(H_{row_j} j = 1...r)$ with the weights contained in the i^{th} row vector of W (W_{row_i}). Therefore, under the viewpoint of the row vectors, H and W are called a *term basis matrix* and a *term coordinate matrix*. Thanks to this interpretation, we propose a clustering-based method for initializing the NMF according to the row (term) vectors. The term viewpoint enables our method to utilize the researches on the term description in the text. By these ways, we overcome the challenge of a clustering-based NMF initialization when defining the starting cluster centroids of the used clustering algorithm. The NMF becomes deterministic and more effective. Sect. 4 presents our clustering-based NMF initialization.

4 A Term Clustering-Based NMF Initialization

4.1 Term Basis Matrix Initialization

Under the term (row) interpretation, a clustering-based NMF initialization becomes clustering the original term vectors $(X_{row_i} i = 1...m)$, and the cluster centroids are then used as the row vectors of the initial term basis matrix $H^{(0)}$. To determine the good starting cluster centroids, it is necessary to select important term vectors from the

original term vectors. In the text classification domain, the notion of important terms implicitly indicates the terms which make a big contribution to the classification. Based on the labeled training set, a wide variety of methods called supervised feature selection (FS) methods [25] is proposed for picking up the most important subset of features (terms) for the purpose of the classification.

Obviously, it is easier to address how a clustering algorithm initializes itself when it is used for initializing the NMF on a Terms-Documents matrix if we change viewing the factors from the column (document) vectors (*Document Clustering-based NMF Initialization* [4, 24, 26]) to the row (term) vector *(Term Clustering-based NMF Initialization*). The Term Clustering-based NMF Initialization is as follows:

- A supervised FS is used to select important terms. They become the first centroids for clustering the term vectors, i.e. the row vectors of Terms-Documents matrix X.
- Thanks to a clustering algorithm, the selected important terms turn into the true cluster centroids. The cluster centroids are pushed into the row vectors of the initial term basis matrix $H^{(0)}$ $(H^{(0)}_{row_i}\, i = 1...r)$.

The supervised FS is known as an offline and relatively low-cost process [10]. It is a major tasks right after the document presentation to completely eliminate noise features without altering the information of the important features. After the FS, an FT combines the remaining important features with each other to form the new and more important features. An FS prior to an FT is to avoid the negative impacts of the noise features on the new features created by the FT as well as to decrease the computational cost of the FT. Therefore, reusing the pre-existing FS results for initializing the NMF does not impose any extra burdens on the NMF computation. For the FS, we aim at our effective supervised term selection named *DtFCFS-BRatTL*. For DtFCFS [18], a term gets a higher score if it makes both the documents in every category become closer and the categories become more separated. Based on the term scores, the BRatTL [17] selects a final term set covering all categories as well as possible.

4.2 Term Coordinate Matrix Initialization

Based on Eq. (2), $W^{(0)}_{ij}$ of the initial term coordinate matrix shows the association between the i^{th} original term (X_{row_i}) and the j^{th} initial term basis (cluster centroid: $H^{(0)}_{row_j}$). For a clustering-based NMF initialization, it is 1 or 0 indicating whether the term X_{row_i} belongs to the cluster $H^{(0)}_{row_j}$ or not. However, this does not show the nature of an FT in resolving the polysemy issue which allows a term to be related to many clusters. To create a better model, we compute $W^{(0)}_{ij}$ $(i = 1...m; j = 1...r)$ based on the *Pointwise Mutual Information* (PMI) [22] between the original term X_{row_i} and initial

term basis $H^{(0)}_{row_j}$. The PMI is one of the effective methods mainly used for measuring the semantic association between two terms as follows:

$$p_{ij} = \frac{f_{ij}}{\sum_{k=1}^{m}\sum_{p=1}^{r} f_{kp}}; p_{i*} = \frac{\sum_{p=1}^{r} f_{ip}}{\sum_{k=1}^{m}\sum_{p=1}^{r} f_{kp}}; p_{*j} = \frac{\sum_{k=1}^{m} f_{kj}}{\sum_{k=1}^{m}\sum_{p=1}^{r} f_{kp}}; W_{ij}^{(0)(pmi)} = \log_2 \frac{p_{ij}}{p_{i*}.p_{*j}} \quad (3)$$

where f_{kp} is co-occurrence value of X_{row_k} and $H^{(0)}_{row_p}$. However, PMI highly values rare terms [22]. [13] indicates that PMI works more effectively when raising p_{*j} to the power of α which is set to 0.75 for the most significant performance as follows:

$$p\_\alpha_{*j} = \frac{\left(\sum_{k=1}^{m} f_{kj}\right)^{\alpha}}{\sum_{p=1}^{r}\left(\sum_{k=1}^{m} f_{kp}\right)^{\alpha}}; W_{ij}^{(0)(pmi\_\alpha)} = \log_2 \frac{p_{ij}}{p_{i*}.p\_\alpha_{*j}} \quad (4)$$

To be compatible with the non-negativity of the NMF, we change from the PMI to Positive PMI (PPMI) [3] ($W_{ij}^{(0)(pmi\_\alpha)} = \max\left(0,\ W_{ij}^{(0)(pmi\_\alpha)}\right)$. [3] points out that the PPMI is better than the PMI and many other methods.

5 Experiment

5.1 Experimental Setup

The experiments are carried out on the 10 top-sized categories of the Newsgroup of the "bydate" split [1], the Reuters of the ModApte split, and the Ohsumed of the Joachims split [11]. Our aim is to investigate the NMF initializations in the study:

- The Document Clustering-based NMF Initialization is implemented with K-Means (KM) (Doc-KM-Cluster-NMFInit [24]); with Fuzzy C-Means (FCM) (Doc-FCM-Cluster-NMFInit [26]); with the Subtractive (SC) (Doc-SC-Cluster-NMFInit [4]).
- The Term Clustering-based NMF Initialization is considered with the KM and the FCM Clustering (Term-KM-Cluster-NMFInit; Term-FCM-Cluster-NMFInit).
- The NNDSVD [2], a well-known factorization-based NMF initialization.
  Figure 1 shows the details in the experimental setup as follows:
- The training set is pre-processed by removing the stop words and word stemming. It is then presented by a Terms-Documents (m x n) with TF-IDF weighting [21]. A supervised FS by the DtFCFS-BRatTL is applied on the training Terms-Documents matrix to remove noise terms as well as decrease the computation cost of the FT.
- After the FS, the L. best terms are selected. The reduced training Terms-Documents matrix (L x n), called X, is taken into the NMF. The NMF is computed by the Multiplicative Update (MU) [12] or Alternate Least Square (ALS) [5] with a NMF rank r. Remember that for a term clustering-based NMF initialization, the FS results are again used. The column vectors of the output H (r x n) are used for building a model using an SVM by SMO [20]. Every new document is converted to a TF-IDF vector t. (L x 1) only based on the terms selected in the FS. Under the semantic

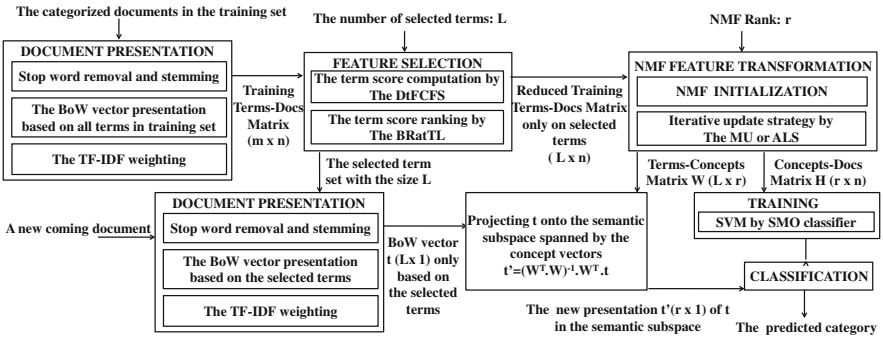

Fig. 1. The experimental setup of the NMF feature transformation for text classification.

space spanned by column vectors of the output W ($L\,x\,r$), its new presentation is computed by the projection $t^{'}$ ($r\,x\,1$) where $t^{'} = (W^T.W)^{-1}.W^T.t$. [24].

Determining the parameter L in an FS and the parameter r in an FT is a big challenge. A common trend in the FS is that the quality of the selected feature (term) set gradually moves towards a saturation point when its size increases. At this point, the rest of features have lower quality, and selecting more features (terms) does not bring high effect. In many our FS researches [15–18] on these experimental datasets, with about the 2000 best selected terms, the classification performance nearly reaches up to the peak. That is the reason why L is set to 2000. Regarding the NMF rank r, it is also the number of clusters in a clustering-based NMF initialization. With the 2000 selected terms for the FT, the maximal number of clusters (rank r) is set up to 600.

5.2 Experimental Result and Discussion

During the iterative process, an NMF aims at minimizing the approximation error. In Fig. 2A, we present the examples about the approximation errors of two NMF algorithms (MU and ALS) under the different initializations when incrementally altering the number of iterations. Firstly, we emphasize the approximation errors at the small iterations, which are heavily affected by the NMF initializations. Noticeably, at small iterations, when the K-Means Clustering (KM) is used, the approximation errors of the NMFs using the term clustering-based initialization (Term-KM-Cluster-NMFInit) are better than those using the document clustering-based initialization (Doc-KM-Cluster-NMFInit). For Fuzzy C-Means Clustering (FCM), this phenomenon is still the same. In comparison with Doc-SC-Cluster-NMFInit and NNDSVD, the NMFs using Term-KM-Cluster-NMFInit obtain more impressive errors in both MU and ALS.

At the increasing iterations, each NMF moves toward its own stability of approximation error, called the convergence point. Similar to the previous researches, at convergence, the approximation errors of NMFs by ALS using the different initializations are nearly equal. That is because an NMF by ALS is less dependent on the initialization. It only needs to initialize W, and H. is computed at the first iteration. In this case, the effectiveness of a NMF initialization is shown through the number of iterations.

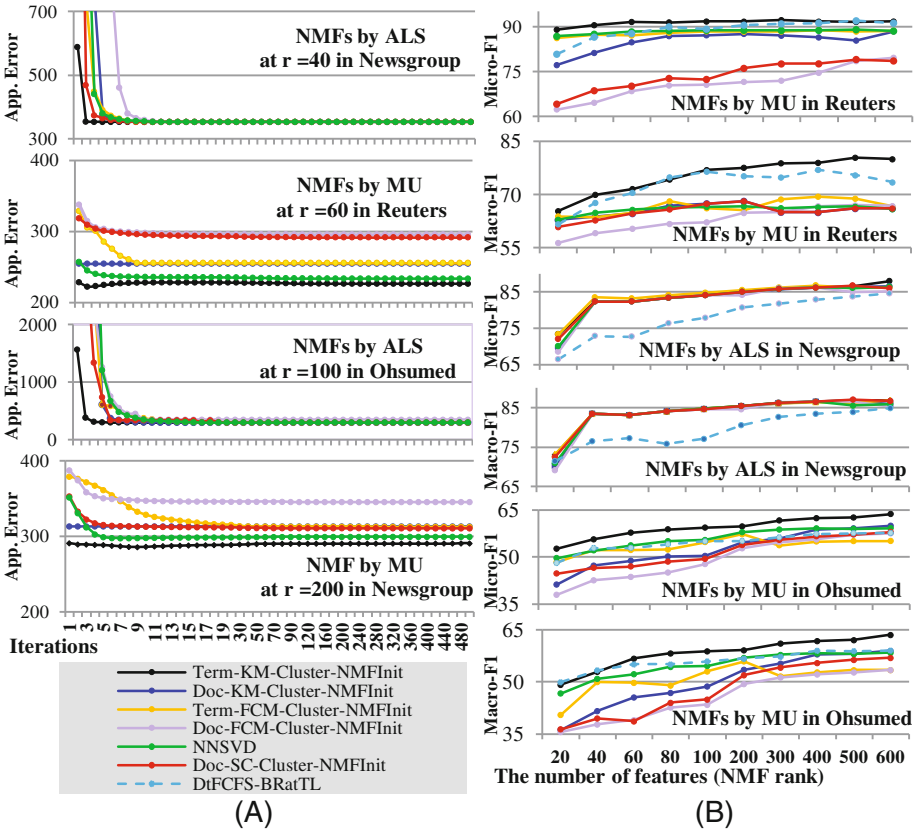

Fig. 2. The NMF in terms of the approximation error (A) and classification performance (B).

As presented in Fig. 2A, NMFs by ALS initialized by the Term-KM-Cluster-NMFInit run quickly toward the convergence. Contrary to the NMFs by ALS, the NMFs by MU have the different errors with the different number of iterations at their convergence points. Concretely, the NMFs by MU using the Doc-KM-Cluster-NMFInit require the fewest number of iterations for convergence. However, at convergence, their approximation errors are larger than those using the Term-KM-Cluster-NMFInit.

In Table 2, the approximation error and the number of iterations of the NMFs are considered under the different initializations at the point satisfying a convergence criterion for more NMF ranks. A popular convergence criterion is that the change of the error in two successive iterations is below 10^{-6} or the number of iterations reaches to 500. Interestingly, what happen here is the same as our analysis on the examples in Fig. 2A. The NMFs by ALS move the equal approximation errors regardless of their different initializations, while the best number of iterations goes to the one initialized by the Term-KM-Cluster-NMFInit. As to the NMFs by MU, the fastest in the race to the convergence is the NMFs by MU initialized by the Doc-KM-Cluster-NMFInit but

Table 2. The approximation error and the number of iterations of the NMF at the convergence point: (1) Term-KM-Cluster-NMFInit; (2) Doc-KM-Cluster-NMFInit; (3) Term-FCM-Cluster-NMFInit; (4) Doc-FCM-Cluster-NMFInit; (5) Doc-SC-Cluster-NMFInit; (6) NNDSVD.

r	20		40		60		80		100		200		300		400		500		600	
	Iter	Err	Iter	Err	Iter	Err	Iter	Err	Iter	Err	Iter	Err	Iter	Err	Iter	Err	Iter	Err	Iter	Err
NMF by MU in the Reuters dataset																				
(1)	358	**256**	412	**239**	286	**226**	314	**224**	208	219	251	**187**	383	218	500	**206**	500	176	**474**	**164**
(2)	**171**	293	**102**	288	**92**	254	**179**	252	**116**	243	**158**	221	**233**	231	**221**	294	500	185	500	185
(3)	500	266	500	248	500	253	389	230	411	224	424	192	500	**167**	500	215	500	132	500	176
(4)	500	317	500	303	500	291	500	285	500	270	500	245	500	241	500	266	500	181	500	193
(5)	500	320	500	305	500	291	500	284	329	276	500	242	500	218	500	296	500	176	500	180
(6)	430	316	271	244	500	235	180	232	396	**217**	500	230	500	186	500	260	500	**161**	500	172
NMF by ALS in the Newsgroup dataset																				
(1)	**162**	346	**27**	354	130	343	**185**	325	500	317	**360**	281	**421**	268	500	233	**335**	222	500	**212**
(2)	222	**345**	31	354	**136**	343	211	325	500	317	476	281	492	280	500	233	500	222	500	213
(3)	259	346	32	354	142	343	174	325	500	317	425	281	456	**255**	500	234	500	223	500	213
(4)	270	346	35	354	146	343	211	325	500	317	500	282	500	282	500	233	500	223	500	213
(5)	275	346	32	354	148	343	210	325	500	317	500	282	500	285	500	233	401	222	500	213
(6)	210	345	32	354	147	343	217	325	500	317	500	281	500	267	500	233	500	223	500	213
NMF by MU in the Ohsumed dataset																				
(1)	89	**333**	253	**322**	170	**314**	145	**306**	500	**300**	500	**279**	500	**248**	**402**	**239**	500	**230**	500	**230**
(2)	**50**	354	**82**	345	**92**	341	**101**	337	**232**	333	**307**	309	**271**	283	500	257	**449**	244	500	244
(3)	259	342	500	334	500	326	500	322	500	317	500	292	500	255	500	249	500	239	500	239
(4)	445	371	500	352	500	348	500	337	500	337	500	311	500	279	500	268	500	253	500	253
(5)	394	355	500	349	83	344	351	340	410	335	500	317	500	292	500	283	468	274	257	274
(6)	438	345	500	335	500	326	207	320	482	314	500	293	500	268	500	261	500	253	500	253

due to the randomization at starting cluster centroids, they can easily fall down a local minimum worse than that of the NMFs by MU initialized by the Term-KM-Cluster-NMFInit. Compared with the other initializations, the Term-KM-Cluster-NMFInit leads the NMFs by MU to a faster convergence and better overall error at convergence. For another used clustering algorithm, i.e. the Fuzzy C-Means Clustering (FCM), the term clustering-based NMF Initialization (Term-FCM-Cluster-NMFInit) is better than the document clustering-based NMF Initialization (Doc-FCM-Cluster-NMFInit). However, it is not more effective than the Term-KM-Cluster-NMFInit.

Next, the NMF initializations are evaluated by the classification performance on the transformed feature set. In order to consider the overall performance of a multi-category classification, two well-known measures, namely Macro-F1 [21] and Micro-F1 [21], are used. Figure 2B shows the Micro-F1 and Macro-F1 results for the NMF FTs. Concerning NMFs by ALS initialized by the different methods, similar to the approximation errors, the Micro-F1 and Maco-F1 results are almost identical. However, as the remarks above on Fig. 2A and Table 2, in order to attain these Micro-F1 and Maco-F1 results, the NMFs by ALS initialized Term-KM-Cluster-NMFInit require the fewest number of iterations. It can be seen from Fig. 2B that the NMFs by MU initialized by the Term-KM-Cluster-NMFInit outperform those initialized by the other methods at most sizes of the feature set in both the Micro-F1 and Macro-F1, while NMFs by MU initialized by the Term-FCM-Cluster-NMFInit produce better Micro-F1 and Macro-F1 than those initialized by the Doc-FCM-Cluster-NMFInit and show competitive and even superior performance to the good methods.

Another emphasis in Fig. 2B is the superiority of the classification on the transformed feature set of the NMFs initialized by the Term-KM-Cluster-NMFInit over that on the non-transformed feature set only simply selected by the DtFCFS-BRatTL. This further confirms the effectiveness of the NMFs initialized by the Term-KM-Cluster-NMFInit. At some sizes of output feature set, the classification performance on the transformed feature set of the NMFs by MU initialized by the document clustering-based methods (Doc-KM-Cluster-NMFInit and Doc-FCM-Cluster-NMFInit) falls even lower than that on the non-transformed feature set selected by the DtFCFS-BRatTL. Therefore, the randomization in starting a document clustering-based NMF initialization has a strong negative influence on the classification performance.

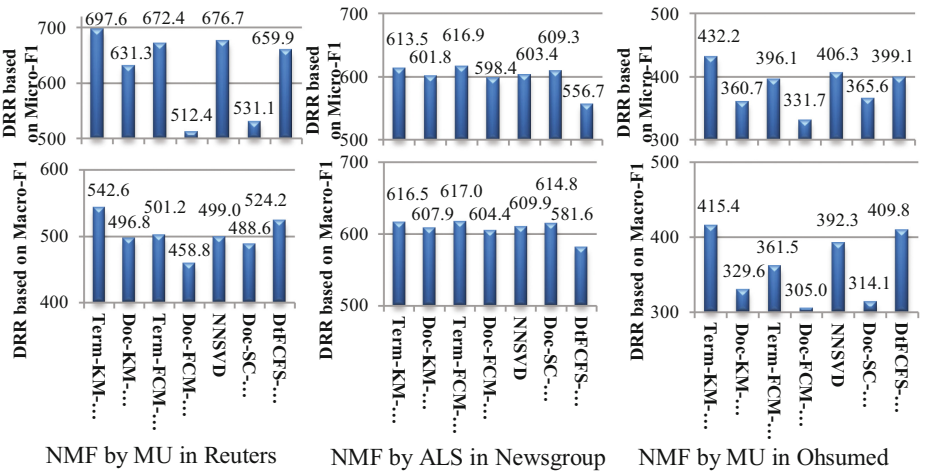

Fig. 3. The NMF performance in terms of the Dimension Reduction Rate (DRR).

Finally, for explicit comparison, we consider Dimension Reduction Rate (DRR) [15, 16, 18] of the NMFs in Fig. 3. The DRR is computed as follows:

$$DRR = \frac{1}{k}\sum_{i=1}^{k}\frac{Dim_N}{Dim_i}R_i \tag{5}$$

where k is the number of experiments; Dim_i is the number of output features (rank) in the i^{th} experiment; R_i is the Micro-F1 or Macro-F1 in the i^{th} experiment; and Dim_N is the maximal number of output features in all experiments. With the Micro-F1 and Macro-F1 results in Fig. 2B, for every of the clustering algorithms, i.e. the K-means and Fuzzy C-Means Clustering, the NMF using the term clustering-based initialization produces more impressive DRRs than that using the document clustering-based initialization. Especially, when the K-means Clustering is used, the term clustering-based NMF initialization (Term-KM-Cluster-NMFInit) shows superior DRRs to the others including NNDSVD, a well-known factorization-based NMF initialization.

Another dominant strength of the term clustering-based NMF initialization is to make the NMF become deterministic. Furthermore, in the experiments on the

document clustering-based NMF initialization, we usually face the issue of empty clusters. In this case, to achieve the best NMF initialization, we must rerun the document clustering process with other random starting cluster centroids. For example, in the Reuters dataset, the document clustering process in the Doc-KM-Cluster-NMFInit is rerun 3 times with 200 clusters; 15 times with 300 clusters; 46 times with 400 clusters; 119 times with 500 clusters; 523 times with 600 clusters. For the Term-KM-Cluster-NMFInit, the issue of empty clusters does not happen even with 600 clusters. This demonstrates the goodness of selecting the starting term cluster centroids by using a supervised FS (DtFCFS-BRatTL) in the term clustering-based NMF initialization.

6 Conclusion

This study may pave the way for further studies on the text FT based on the NMF not only at the NMF initialization stage but also the other NMF stages by interpreting the NMF matrix factors according to their row (term) vectors. Under the viewpoint of term vectors, it is possible to exploiting the researches on term description in the text to further improve the NMF FT. For instance, in this paper, we utilize the DtFCFS-BRatTL, which is a recent supervised term selection of ours, and the PPMI, which is an effective semantic term association, to propose a new clustering-based method for initializing the NMF matrix factors according to the row (term) vectors instead of the column (document) vectors. And it is called a term clustering-based NMF initialization. This facilitates settling how a clustering algorithm defines the starting cluster centroids when it is used for the NMF initialization. Therefore, one of the dominant strengths of the term clustering-based NMF initialization is to make the NMF become deterministic. We investigate the performance of the document clustering-based NMF initializations and the proposed term clustering-based NMF initializations. The results show that the NMFs by ALS obtain the nearly equal approximation errors, classification performance, and dimension reduction rate regardless of their different initializations. However, when the K-means clustering is used, the term clustering-based NMF initializations leads the NMFs by ALS to a faster convergence. For the NMFs by MU, the term clustering-based NMF initialization is better than that using the document clustering-based initialization in terms of the approximation error, the classification performance, and the dimension reduction rate. Especially, with K-Means Clustering, the term clustering-based NMF initialization is superior to the others.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
2. Boutsidis, C., Gallopoulos, E.: SVD based initialization: a head start for nonnegative matrix factorization. Patt. Recogn. **41**(4), 1350–1362 (2008)
3. Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: a computational study. Behav. Res. Methods **39**(3), 510–526 (2007)
4. Casalino, G., Del Buono, N., Mencar, C.: Subtractive clustering for seeding non-negative matrix factorizations. Inf. Sci. **257**, 369–387 (2014)

5. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley, Chichester (2009)
6. Correa, R.F., Ludermir, T.B.: Improving self-organization of document collections by semantic mapping. Neurocomputing **70**(1), 62–69 (2006)
7. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391 (1990)
8. Golub, G.H., Van Loan, C.F.: Matrix Computations, vol. 3. JHU Press, Baltimore (2012)
9. Hosseini-Asl, E., Zurada, Jacek M.: Nonnegative matrix factorization for document clustering: a survey. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, Lotfi A., Zurada, Jacek M. (eds.) ICAISC 2014. LNCS, vol. 8468, pp. 726–737. Springer, Cham (2014). doi:10.1007/978-3-319-07176-3_63
10. Janecek, A., Gansterer, W.N., Demel, M., Ecker, G.: On the relationship between feature selection and classification accuracy. In: FSDM, pp. 90–105 (2008)
11. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer, Heidelberg (1998). pp. 137–142
12. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562 (2001)
13. Levy, O., Gold, Y.: Improving distributional similarity with lessons learned from word embeddings. Trans. Comput. Linguist. Assoc. **3**, 211–225 (2015)
14. Liu, H., Motoda, H. (Eds.): Feature Extraction, Construction and Selection: A Data Mining Perspective. Springer, New York (1998)
15. Nam, L.N.H., Quoc, H.B.: A comprehensive filter feature selection for improving document classification. In: Proceedings of 29th Pacific Asia Conference on Language, Information and Computation 2015, pp. 169–177 (2015)
16. Nam, L.N.H., Quoc, H.B.: A combined approach for filter feature selection in document classification. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 317–324. IEEE (2015)
17. Nam, L.N.H., Quoc, H.B.: The ranking methods in the filter feature selection process for text categorization system. In: Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS 2016) (Paper 159) (2016)
18. Nam, L.N.H., Quoc, H.B.: The hybrid filter feature selection methods for improving high-dimensional text categorization. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **25** (02), 235–265 (2017)
19. Pinheiro, R.H., Cavalcanti, G.D.: Data-driven global-ranking local feature selection methods for text categorization. Expert Syst. Appl. **42**(4), 1941–1949 (2015)
20. Platt, J.C.: 12 fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods, pp. 185–208 (1999)
21. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**(1), 1–47 (2002)
22. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**(1), 141–188 (2010)
23. Wang, Y.X., Zhang, Y.J.: Nonnegative matrix factorization: a comprehensive review. IEEE Trans. Knowl. Data Eng. **25**(6), 1336–1353 (2013)
24. Xue, Y., Tong, C.S., Chen, Y.: Clustering-based initialization for non-negative matrix factorization. Appl. Math. Comput. **205**(2), 525–536 (2008)
25. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML, vol. 97, pp. 412–420, July 1997
26. Zheng, Z., Yang, J., Zhu, Y.: Initialization enhancer for non-negative matrix factorization. Eng. Appl. Artif. Intell. **20**(1), 101–110 (2007)