2010 IEEE International Conference on Data Mining

Weighted Feature Subset Non-Negative Matrix Factorization and its Applications to Document Understanding

Dingding Wang Tao Li
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
Email: {dwang003,taoli}@cs.fiu.edu

Chris Ding
Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX, USA
Email: chqding@uta.edu

***Abstract*—Keyword (Feature) selection enhances and improves many Information Retrieval (IR) tasks such as document categorization, automatic topic discovery, etc. The problem of keyword selection is usually solved using supervised algorithms. In this paper, we propose an unsupervised approach that combines keyword selection and document clustering (topic discovery) together. The proposed approach extends non-negative matrix factorization (NMF) by incorporating a weight matrix to indicate the importance of the keywords. The proposed approach is further extended to a weighted version in which each document is also assigned a weight to assess its importance in the cluster. This work considers both theoretical and empirical weighted feature subset selection for NMF and draws the connection between unsupervised feature selection and data clustering. We apply our proposed approaches to various document understanding tasks including document clustering, summarization, and visualization. Experimental results demonstrate the effectiveness of our approach for these tasks.**

***Keywords*-Non-negative matrix factorization; feature selection; weighted feature subset non-negative matrix factorization.**

I. Introduction

Recently, many research efforts have been reported on developing efficient and effective techniques for analyzing large document collections. Among these efforts, nonnegative matrix factorization (NMF) has been shown to be useful for different document understanding problems, e.g., document clustering [40] and summarization [38]. The success of NMF is largely due to the newly discovered ability of NMF to solve challenging data mining and machine learning problems. In particular, NMF with the sum of squared error cost function is equivalent to a relaxed K-means clustering, the most widely used unsupervised learning algorithm [8]. In addition, NMF with the I-divergence cost function is equivalent to probabilistic latent semantic indexing (PLSI) [22], another unsupervised learning method popularly used in text analysis [10], [14]. Furthermore, NMF is able to model widely varying data distributions and can do both hard and soft clustering simultaneously. Several variants of NMF with different forms of factorization and regularization have also been developed and applied to many document analysis tasks [11], [18], [38], [39].

Although NMF and its variants have shown their effectiveness in these tasks, they usually perform data clustering on all feature space. As we know, keyword (feature) selection can enhance and improve many document applications such as document categorization and automatic topic discovery. However most of existing keyword selection techniques are designed for supervised classification problems.

In this paper, we extends NMF to solve a novel problem of clustering with double labeling of important features and data points, which means that each data point is marked as belonging to one of the groups, and each feature and data point are also weighted to assess their importance respectively.

In particular, we first extends NMF to feature subset NMF which combines keyword selection and document clustering (topic discovery) together. The proposed extension incorporates a weight matrix to indicate the importance of the keywords. It considers both theoretically and empirically feature subset selection for NMF and draws the connection between unsupervised feature selection and data clustering. The selected keywords are discriminant for different topics in a global perspective, unlike those obtained in co-clustering, which typically associate with one cluster strongly and are absent from other clusters. Also, the selected keywords are not linear combinations of words like those obtained in Latent Semantic Indexing (LSI) [17]: our selected words provide clear semantic meanings of the key features while LSI features combine different words together and are not easy to interpret.

We further extend feature subset NMF into a weighted version which assumes documents (data points) contribute differently to the clustering process, i.e., some documents are tightly related to certain topics, while some can be considered as outliers. Finally, we apply the proposed approaches in document understanding problems such as document clustering, summarization, and visualization. The comprehensive experiments demonstrate the effectiveness of our approaches.

The rest of the paper is organized as follows. Section II discusses the related work in NMF framework and various

1550-4786/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDM.2010.47

document understanding tasks. In Section III, we derive a generic theorem on the NMF algorithm. Section IV and Section V propose our (weighted) feature subsect NMF. An illustrative example is shown in Section VI, and comprehensive experiments on document clustering, summarization, and visualization are conducted in Section VII. Section VIII concludes.

II. Related Work

A. NMF Framework

NMF has been shown to be very useful for data clustering. Lee and Seung [24] proposed the NMF problem and showed that the NMF problem could be solved by a multiplicative update algorithm. In general, the NMF algorithm attempts to find the subspaces in which the majority of the data points lie. Let the input data matrix $X = (x_1, \ldots, x_n)$ contain the collection of n nonnegative data column vectors. The problem of NMF aims to factorize X into two nonnegative matrices,

$$X \approx FG^T,$$

where $X \in \mathbb{R}_+^{p\times n}$, $F \in \mathbb{R}_+^{p\times k}$, and $G \in \mathbb{R}_+^{n\times k}$.

Similarly, there are other matrix factorizations which differ with standard NMF by the restrictions on the matrix factors and forms: We list them as follows.

Convex-NMF:	$X \approx XWG^T$
Tri-Factorization:	$X \approx FSG^T$
WFS-NMF:	$min\|\|X - FG^T\|\|_W^2$

Note that WFS-NMF is our proposed algorithm which extends NMF by incorporating a weight matrix to indicate the importance of the keywords and data points. The detail of the algorithm will be discussed in the following sections. A preliminary study of feature subset NMF which only considers the importance of keywords was presented as a two-page poster [37].

The relations between NMF and some of the other matrix factorization and clustering algorithms have been studied in [25]. In general, (1) Orthogonal NMF is equivalent to K-means clustering; (2) G-orthogonal NMF, semi-NMF and convex-NMF are identical to relaxed K-means clustering; (3) Tri-factorization with explicit orthogonality constraints can be transformed into 2-factor NMF; (4) PLSI [22] (which is further developed into a more comprehensive Latent Dirichlet Allocation (LDA) model [1]) solves the problem of NMF with Kullback-Leibler divergence; (5) our proposed WFS-NMF combines clustering with double labeling of important features and samples by assigning different weights to each row and column based on the weight matrix.

B. Document Understanding Applications

There exist various document understanding applications in IR community. Here, we briefly review some popular tasks including document clustering, document summarization, and visualization. In this paper, we also apply our proposed approaches to these three applications.

Document Clustering.: The problem of document clustering has been extensively studied. Given a collection of documents, document clustering partitions them into different groups (called clusters) so that similar documents belong to the same group while the documents in different clusters are dissimilar. The problem of document clustering has been extensively studied. Traditional clustering techniques such as hierarchical and partitioning methods have been used in clustering documents (e.g. hierarchical agglomerative clustering (HAC) [12] and K-means clustering [20]). Model-based clustering methods such as PLSI and the more comprehensive LDA have also been successfully applied to document clustering [22], [1]. Recently, matrix and graph based clustering algorithms have emerged as promising clustering approaches [39], and two representative examples of which are spectral clustering [34] and non-negative matrix factorization (NMF) [24], [40]. Co-clustering algorithms are then proposed which aim at clustering document and term simultaneously by making use of the dual relationship information [5], [7], [43]. Subspace clustering algorithms have also been developed for discovering low-dimensional clusters in high-dimension document space [26], [23].

Multi-Document Summarization.: Multi-document summarization aims to generate a short summary for a collection of documents reflecting the major or query-relevant information. Existing summarization methods usually rank the sentences in the documents according to their salient scores calculated by a set of predefined liguistic features, such as term frequency-inverse sentence frequency (TF-ISF) [28], sentence or term position [41], and number of keywords [41]. Gong et al. [16] propose a generic method using latent semantic analysis (LSA) to select sentences with high ranking for summarization. Goldstein et al. [15] propose a maximal marginal relevance (MMR) method to summarize documents based on the cosine similarity between a query and a sentence and also the sentence and previously selected sentences. Other approaches include NMF based summarization [30], Conditional Random Field (CRF) based summarization [33], and hidden Markov model (HMM) based method [4]. In addition, graph-ranking based approaches have been proposed to summarize documents using the sentence relationship [13], the idea of which is similar to PageRank.

Document Visualization.: Document visualization focuses on displaying document relationships using various presentation techniques, which helps users to understand and navigate information easily. Some techniques have been developed to map the document collection into multivariate space. Typical systems for document visualization include the Galaxy of News [32], Jigsaw [35], and ThemeRiver [21].

In this paper, we extend the NMF model to allow unsu-

pervised feature selection and data clustering and ranking to be conducted simultaneously. We apply the proposed approaches in three document understanding applications to demonstrate the effectiveness of the approaches for improving document understanding.

III. A GENERIC THEOREM ON NMF ALGORITHM

In this paper, we will derive several algorithms for NMF problems. Here we first provide a generic theorem on the NMF algorithm. We will use this results repeatedly later.

For the following optimization problem

$$\min_{H\geq 0} \; J(H) = \text{Tr}[-2R^T H + H^T PHQ], \tag{1}$$

where $P, Q, H \geq 0$ are constant matrices, the optimal solution for H is given by the following updating algorithm

$$H_{ik} \leftarrow H_{ik}\frac{R_{ik}}{(PHQ)_{ik}}. \tag{2}$$

Theorem 1. If the algorithm converges, the converged solution satisfies the KKT condition.

Proof. We minimize the Lagrangian function

$$\mathcal{L}(H) = \text{Tr}[-2R^T H + H^T PHQ - 2\beta H],$$

where $\lambda = (\lambda_{ik})$ is the Lagrangian multiplier to enforce $H_{ik} \geq 0$. Setting

$$\frac{\partial J}{\partial H_{ik}} = (-2R + 2PHQ - 2\beta)_{ik} = 0,$$

the KKT complementarity slackness condition $\beta_{ik}H_{ik} = 0$ becomes

$$(-R + PHQ)_{ik}H_{ik} = 0. \tag{3}$$

Now, when iterative solution of H convergees, it satisfies

$$H_{ik} = H_{ik}\frac{R_{ik}}{(PHQ)_{ik}}. \tag{4}$$

One can see Eq.(4) is identical to Eq.(3) either $H_{ik} = 0$ or not. This proves that the converged solution satisfies KKT condition. □

Theorem 2. The updating algorithm of Eq.(2) converges.

Proof. We use the auxiliary function approach [24]. A function $Z(H, H')$ is called an auxiliary function of $J(H)$ if it satisfies

$$Z(H, H') \geq J(H), \quad Z(H, H) = J(H), \tag{5}$$

for any H, H'. Define

$$H^{(t+1)} = \arg\min_{H} Z(H, H^{(t)}), \tag{6}$$

where we note that we require the global minimum. By construction, we have $J(H^{(t)}) = Z(H^{(t)}, H^{(t)}) \geq Z(H^{(t+1)}, H^{(t)}) \geq J(H^{(t+1)})$. Thus $J(H^{(t)})$ is monotone decreasing (non-increasing). The key is to find (1) appropriate $Z(H, H')$ and (2) its global minimum.

Using the following matrix inequality

$$\text{Tr}(H^T PHQ) \geq \sum_{ik} \frac{(PH'Q)_{ik}H_{ik}^2}{H'_{ik}},$$

where $H, P, Q \geq 0$ and $P = P^T, Q = Q^T$, we can see that

$$Z(H, H') = \sum_{ik} R_{ik}H_{ik} + \sum_{ik} \frac{(PH'Q)_{ik}H_{ik}^2}{H'_{ik}}$$

is an auxiliary function of $J(H)$ of Eq.(1). Now we solve Eq.(6) by identifying $H^{(t+1)} = H$ and $H^{(t)} = H'$. Setting

$$\frac{\partial Z}{\partial H_{ik}} = -2R_{ik} + 2\frac{(PH'Q)_{ik}H_{ik}}{H'_{ik}} = 0, \tag{7}$$

we obtain

$$H_{ik} = H'_{ik}\frac{R_{ik}}{(PH'Q)_{ik}}. \tag{8}$$

The second derivatives are

$$\frac{\partial^2 Z}{\partial H_{ik}\partial H_{jl}} = 2\frac{(PH'Q)_{ik}}{H'_{ik}}\delta_{ij}\delta_{kl},$$

which is a semi-positive definite matrix, ensuring the local optima of Eq.(8) obtained from Eq.(7) is the global minima for solving Eq.(6). Thus updating H using Eq.(8) will decrease $J(H)$. One can see Eq.(8) is identical to Eq.(2). □

IV. FEATURE SUBSET NMF (FS-NMF)

A. Objective

Let $X = \{x_1, \cdots, x_n)$ contains n documents with m keywords (features). In general, NMF factorizes the input nonnegative data matrix X into two nonnegative matrices,

$$X \approx FG^T,$$

where $G \in \mathbb{R}_+^{n\times k}$ is the cluster indicator matrix for clustering columns of X and $F = (f_1, \cdots, f_k) \in \mathbb{R}_+^{m\times k}$ contains k cluster centroids.

In this paper, we propose a new objective to simultaneously factorize X and rank the features in X as follows:

$$\min_{W\geq 0, F\geq 0, G\geq 0} ||X - FG^T||_W^2, s.t. \sum_j W_j^\alpha = 1, \tag{9}$$

where $W \in \mathbb{R}_+^{m\times m}$ which is a diagonal matrix indicating the weights of the rows (keywords or features) in X, and α is a parameter (set to 0.7 empirically).

B. Optimization

Minimizing Eq.(9) with respect to W, F, and G, has a closed-form solution. We will optimize the objective with respect to one variable while fixing the other variables. This procedure repeats until convergence.

1) Computation of W: Optimizing Eq.(9) with respect to W is equivalent to optimizing

$$J_1 = \sum_i W_i u_i - \lambda(\sum_i W_i^\alpha - 1),\ u_i = \sum_j (X - FG^T)_{ij}^2.$$

Now, from the KKT condition $\frac{\partial J_1}{\partial W_i} W_i = (u_i - \lambda\alpha W_i^{\alpha-1}) W_i = 0$, we obtain the following updating formula

$$W_i = \left[\frac{1}{\sum_i u_i^{\frac{\alpha}{\alpha-1}}}\right]^{\frac{1}{\alpha}} u_i^{\frac{1}{\alpha-1}}. \tag{10}$$

2) Computation of G: Optimizing Eq.(9) with respect to G is equivalent to optimizing

$$J_2(G) = \mathrm{Tr}(X^T W^T X - 2GF^T W^T X + F^T W^T FG^T G).$$

Using the generic algorithm in Section III, we obtain the following updating formula

$$G_{ik} \leftarrow G_{ik} \frac{(X^T WF)_{ik}}{(GF^T WF)_{ik}}. \tag{11}$$

3) Computation of F: Optimizing Eq.(9) with respect to F is equivalent to optimizing

$$J_3(F) = \mathrm{Tr}[WXX^T - 2WXGF^T + WFG^T GF].$$

Using the generic algorithm in Section III, we obtain the following updating formula

$$F_{ik} \leftarrow F_{ik} \frac{(WXG)_{ik}}{(WFG^T G)_{ik}}. \tag{12}$$

4) Algorithm Procedure: The detail procedure of FS-NMF is listed as Algorithm 1.

Algorithm 1 FS-NMF Algorithm Description

Input:	X	: word-document matrix
	K	: the number of clusters
Output:	F	: word cluster indicator matrix
	G	: document cluster indicator matrix
	W	: word weights matrix

1: Initialize $W = I$ and initialize (F, G) as the output of standard NMF
2: **repeat**
3: Update W by $W_i = \left[\frac{1}{\sum_i u_i^{\frac{\alpha}{\alpha-1}}}\right]^{\frac{1}{\alpha}} u_i^{\frac{1}{\alpha-1}}$, where $u_i = \sum_j (X - FG^T)_{ij}^2$;
4: Update G by $G_{ik} \leftarrow G_{ik} \frac{(X^T WF)_{ik}}{(GF^T WF)_{ik}}$;
5: Update F by $F_{ik} \leftarrow F_{ik} \frac{(WXG)_{ik}}{(WFG^T G)_{ik}}$;
6: **until** converges.

V. Weighted Feature Subset NMF (WFS-NMF)

In Section IV, different weights are assigned to the term features indicating the importance of the keywords, however all the documents are treated equally. This assumption does no longer hold in case that different documents are created with different importance. Thus, we extend our algorithm to a weighted version in which each document is also assigned a weight.

Similar to Eq.(9), the objective of weighted FS-NMF can be written as:

$$\min_{W\geq 0, F\geq 0, G\geq 0} ||X - FG^T||_W^2,$$

where we set $W_{ij} = a_i b_j$. This becomes

$$\min_{W\geq 0, F\geq 0, G\geq 0} (X - FG^T)_{ij}^2 a_i b_j,$$
$$s.t. \sum_i a_i^\alpha = 1, \sum_j b_j^\beta = 1, \tag{13}$$

where α, β are two parameters with $0 < \alpha < 1, 0 < \beta < 1$.

A. Optimization

1) Computation of W: Since $W = \mathbf{a}\mathbf{b}^T$, we optimize $\mathbf{a} = (a_1, \cdots, a_m)$ first. Optimizing Eq.(13) with respect to $\mathbf{a}$ is equivalent to optimizing

$$J_a = \sum_i u_i a_i - \lambda(\sum_i a_i^\alpha - 1),\ u_i = \sum_j (X - FG^T)_{ij}^2 b_j.$$

This optimization has been analyzed in Section IV-B. The optimal solution for $\mathbf{a}$ is given by

$$a_i = \left[\frac{1}{\sum_i u_i^{\frac{\alpha}{\alpha-1}}}\right]^{\frac{1}{\alpha}} u_i^{\frac{1}{\alpha-1}}. \tag{14}$$

We now optimize the objective Eq.(13) with respect to $\mathbf{b} = (b_1, \cdots, b_n)$ which is equivalent to optimizing

$$J_b = \sum_j v_j b_j - \lambda(\sum_j b_j^\beta - 1),\ v_j = \sum_i (X - FG^T)_{ij}^2 a_i.$$

The optimal solution for $\mathbf{b}$ is given by

$$b_j = \left[\frac{1}{\sum_j v_j^{\frac{\beta}{\beta-1}}}\right]^{\frac{1}{\beta}} v_j^{\frac{1}{\beta-1}}. \tag{15}$$

2) Computation of F: Let $A = \mathrm{diag}(a_1, a_2, \ldots, a_m)$ and $B = \mathrm{diag}(b_1, b_2, \ldots, b_n)$. Optimizing Eq.(13) with respect to F is equivalent to optimizing

$$J_4(F) = \sum_{ij} [\sqrt{a_i}(X - FG^T)_{ij} \sqrt{b_j}]^2$$
$$= \|A^{\frac{1}{2}}(X - FG^T)B^{\frac{1}{2}}\|^2$$
$$= \mathrm{Tr}(X^T AXB - 2G^T BX^T AF + F^T AFG^T BG). \tag{16}$$

Using the generic algorithm of Section III, we obtain

$$F_{ik} \leftarrow F_{ik} \frac{(AXBG)_{ik}}{(AFG^T BG)_{ik}}. \tag{17}$$

3) Computation of G: Using Eq.(16), the objective for G is

$$J_5(G) = \mathrm{Tr}(X^T AXB - 2G^T BX^T AF + G^T BGF^T AF). \quad (18)$$

Using the generic algorithm of Section III, we obtain

$$G_{jk} \leftarrow G_{jk} \frac{(BX^T AF)_{jk}}{(BGF^T AF)_{jk}}. \quad (19)$$

B. Algorithm Procedure

The detail procedure of WFS-NMF is listed as Algorithm 2.

Algorithm 2 WFS-NMF Algorithm Description

Input: X : word-document matrix
K : the number of clusters
Output: F : word cluster indicator matrix
G : document cluster indicator matrix
W : word and document weights matrix

1: Initialize $W = I$ and initialize (F, G) as the output of standard NMF
2: **repeat**
3: Update W by $W_{ij} = a_i b_j$,
$a_i = \left[\frac{1}{\sum_i u_i^{\frac{\alpha}{\alpha-1}}}\right]^{\frac{1}{\alpha}} u_i^{\frac{1}{\alpha-1}}$, $b_j = \left[\frac{1}{\sum_j v_j^{\frac{\beta}{\beta-1}}}\right]^{\frac{1}{\beta}} v_j^{\frac{1}{\beta-1}}$,
where $u_i = \sum_j (X - FG^T)_{ij}^2$
and $v_j = \sum_i (X - FG^T)_{ij}^2$;
4: Update G by $G_{jk} \leftarrow G_{jk} \frac{(BX^T AF)_{jk}}{(BGF^T AF)_{jk}}$;
5: Update F by $F_{ik} \leftarrow F_{ik} \frac{(AXBF^T)_{jk}}{(AFGBG^T)_{ik}}$;
6: **until** converges.

VI. An Illustrative Example

In this section, we use a simple example to illustrate the process of weighting the keywords and data points using the proposed WFS-NMF algorithm.

An example dataset with six system log messages is presented in Table I, which is a subset of the Log data described in Section VII-A1. The six sample messages belong to two different clusters: "start" and "create".

	Start
S1	User profile application version 1.0 started successfully.
S2	Database application version 1.1 starts.
S3	Start application version 2.0 for temporary services.
	Create
S4	Can not create temporary ser vices for the Oracle engine.
S5	Can not create temporary services on the files.
S6	Create application version 2.0 for temporary services.

Table I
An example dataset with two clusters .

In the data pre-processing step, the stop words and the words which only appear once are removed, and also stemming is performed. The following term-message matrix is obtained after pre-processing,

$$X = \begin{pmatrix} & S1 & S2 & S3 & S4 & S5 & S6 \\ start & 1 & 1 & 1 & 0 & 0 & 0 \\ application & 1 & 1 & 1 & 0 & 0 & 1 \\ version & 1 & 1 & 1 & 0 & 0 & 1 \\ create & 0 & 0 & 0 & 1 & 1 & 1 \\ temporary & 0 & 0 & 1 & 1 & 1 & 1 \\ service & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

After the computation by WFS-NMF, the weights for the terms are

$$a = \begin{pmatrix} start & 0.1841 \\ application & 0.1581 \\ version & 0.1581 \\ create & 0.1970 \\ temporary & 0.1651 \\ service & 0.1651 \end{pmatrix}.$$

Thus the most important two keywords are "start" and "create", which is consistent with our perspective. Similarly, the weights for the messages are

$$b = \begin{pmatrix} S1 & 0.1743 \\ S2 & 0.1743 \\ S3 & 0.1512 \\ S4 & 0.1839 \\ S5 & 0.1839 \\ S6 & 0.1600 \end{pmatrix}.$$

Then we know S3 and S6 are not important words in discriminating the two clusters as they have the lowest weights. From the example, we clearly observe that the proposed approaches can discover key features and samples.

VII. Experiments

A. Document Clustering

First of all, we examine the clustering performance of FS-NMF and W-FS-NMF using four text datasets as described in Section VII-A1, and compare the results with seven widely used document clustering methods as described in Section VII-A2.

Datasets	# Samples	# Dimensions	# Class
CSTR	475	1000	4
Log	1367	200	8
Reuters	2900	1000	10
WebACE	2340	1000	20

Table II
Dataset Descriptions.

1) Data Sets: Table II summarizes the characteristics of the datasets used in the experimetns. Detailed descriptions of the data sets are as follows.

- **CSTR**. This is the dataset of the abstracts of technical reports (TRs) published in the Department of Computer Science at University of Rochester from 1991 to 2002. The dataset contained 476 abstracts, which were divided into four research areas: Natural

Language Processing(NLP), Robotics/Vision, Systems, and Theory.

- **Log**. This dataset contains 1367 log text messages collecctd from several diff erent machines at Florida International University with different operating systems using logdump2td (an NT data collection tool). There are 9 categories of these messages, i.e., configuration, connection, create, depende ncy, other, report, request, start, and stop.
- **Reuters**. The Reuters-21578 Text Categorization Test collection contains documents collected from the Reuters newswire in 1987. It is a standard text categorization benchmark and contains 135 categories. In our experiments, we use a subset of the data collection which includes the 10 most frequent categories among the 135 topics and we call it Reuters-top 10.
- **WebAce**. This is from WebACE project and has been used for document clustering [2], [19]. The dataset contains 2340 documents consisting news articles from Reuters new service via the Web in October 1997. These documents are divided into 20 classes. Newsgroups. The 20 newsgroups dataset contains approximately 20,000 articles evenly divided among 20 Usenet newsgroups. The raw text size is 26MB.

To pre-process the datasets, we remove the stop words using a standard stop list, all HTML tags are skipped and all header fields except subject and organization of the posted articles are ignored. In all our experiments, we first select the top 1000 words by mutual information with class labels. The feature selection is done with the rainbow package [29].

2) Implemented Baselines: We compare the clustering performance of FS-NMF and W-FS-NMF with the following most widely used document clustering methods. (1) ***K-means***: Stardard K-means algorithm; (2) ***PCA-Km***: PCA is firstly applied to reduce the data dimension followed by the K-means clustering; (3) ***LDA-Km*** [9]: an adaptive subspace clustering algorithm by integrating linear discriminant analysis (LDA) and K-means clustering into a coherent process; (4) ***ECC***: Euclidean co-clustering [3]; (5) ***MSRC***: minimum squared residueco clustering [3]; (6) **NMF**: Nonnegative matrix factorization [40]; (7) ***TNMF***: Tri-factor matrix factorization [11]; (8) ***Ncut***: Spectral Clustering with Normalized Cuts [42].

In these implemented baselines, (a) the K-means algorithm is one of the most widely used standard clustering algorithm; (b) LDA-Km and PCA-Km are two subspace clustering algorithms which identify clusters existing in the subspaces of the original data space; (c) Spectral Clustering with Normalized Cuts (Ncut) is also implemented since it has been shown that that weighted Kernel K-means is equivalent to the normalized cut [6]; (d) both ECC and MSRC are document co-clustering algorithms that are able to find blocks in a rectangle document-term matrix. Co-clustering algorithms generally perform implicit dimension reduction during clustering process. NMF has been shown to be effective in document clustering [40], and our methods are both based on the NMF framework.

3) Evaluation Measures: To measure the clustering performance, we use accuracy and normalized mutual information as our performance measures. Accuracy discovers the one-to-one relationship between clusters and classes and measures the extent to which each cluster contained data points from the corresponding class. It sums up the whole matching degree between all pair class-clusters. Its value is between $[0, 1]$. Accuracy can be represented as:

$$ACC = \max(\sum_{C_i, L_j} T(C_i, L_j))/N, \tag{20}$$

where C_i denotes the i-th cluster, and L_j is the j-th class. $T(C_i, L_j)$ is the number of entities which belong to class j are assigned to cluster i. Accuracy computes the maximum sum of $T(C_i, L_j)$ for all pairs of clusters and classes, and these pairs have no overlaps. Generally, the greater accuracy means the better clustering performance.

Normalized mutual information (NMI) is another widely used performance evaluation measure for determining the quality of clusters [36]. For two random variables X and Y, the NMI is defined as

$$\text{NMI}(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}}, \tag{21}$$

where $I(X, Y)$ is the mutual information between X and Y, and $H(X)$ and $H(Y)$ are the entropies of X and Y, respectively. Clearly, $\text{NMI}(X, X) = 1$ and this is the maximum possible value of NMI. Given a clustering result, NMI in Eq.(21) is estimated as follows:

$$\text{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} \log(\frac{n \cdot n_{ij}}{n_i \cdot \hat{n}_j})}{\sqrt{(\sum_{i=1}^{k} n_i \log \frac{n_i}{n})(\sum_{j=1}^{k} \hat{n}_j \log \frac{\hat{n}_j}{n})}} \tag{22}$$

where n_i denotes the number of data points contained in the cluster $C_i (1 \leq i \leq k)$, $\hat{n}_j$ is the number of data points belonging to the j-th class ($1 \leq j \leq k$), and n_{ij} denotes the number of data points that are in the intersection between the cluster C_i and the j-th class. In general, the larger the NMI value, the better the clustering quality.

4) Clustering Results: Table III and Table IV show the accuracy and NMI evaluation results on the text datasets. From the experimental comparisons, we observe that:

- On most datasets, subspace clustering algorithms (especially LDA-Km) outperform the standard K-means algorithm due to the pre-processing by LDA or PCA.
- Co-clustering algorithms (ECC and MSRC) generally outperform K-means since they are performing implicit dimension reduction during the clustering process.
- NMF outperforms K-means significantly since NMF can model widely varying data distributions due to the flexibility of matrix factorization as compared to

	WebACE	Log	Reuters	CSTR
K-means	0.4081	0.6979	0.4360	0.5210
PCA-Km	0.4432	0.6562	0.3925	0.5630
LDA-Km	0.4774	0.7198	**0.5142**	0.5630
ECC	0.4081	0.7228	0.4968	0.5210
MSRC	0.4432	0.5655	0.4516	0.5630
NMF	0.4774	0.7608	0.4047	0.5630
TNMF	0.4996	0.7527	0.4682	0.6008
Ncut	0.4513	0.7574	0.4890	0.5435
FS-NMF	**0.5577**	0.7715	0.4697	0.6996
WFS-NMF	0.5371	**0.7732**	0.4992	**0.7899**

Table III
CLUSTERING ACCURACY.

	WebACE	Log	Reuters	CSTR
K-means	0.3943	0.4156	0.4033	0.3378
PCA-Km	0.4162	0.5829	0.4362	0.0.3749
LDA-Km	0.4526	0.6542	0.4362	0.3765
ECC	0.4355	0.6028	0.4524	0.4365
MSRC	0.4280	0.4738	0.4251	0.3374
NMF	0.4758	0.7306	0.4391	0.4222
TNMF	0.4799	0.7341	0.4725	0.5138
Ncut	0.4925	0.6386	0.4645	0.4635
FS-NMF	0.4965	**0.7403**	0.4467	0.5636
WFS-NMF	**0.5209**	0.7367	**0.4926**	**0.5817**

Table IV
CLUSTERING NMI RESULTS.

the rigid spherical clusters that the K-means clustering objective function attempts to capture [8].

- TNMF provides a good framework for simultaneously clustering the rows and columns of the input documents. Hence TNMF generally outperforms NMF.
- The results of spectral clustering (Ncut) is better than K-means. Note that spectral clustering can be viewed as a weighted version of Kernel K-means and hence it is able to discover arbitrarily shaped clusters. The experimental results of Ncut is similar to those of NMF. Note that it has also been that NMF is equivalent to spectral clustering [8].
- The proposed FS-NMF and WFS-NMF extend the NMF model and provide a good framework for weighting different terms and documents. Hence both of them generally outperform NMF and TNMF on the datasets. And in the meanwhile, important term features can be discovered by our algorithms.
- As the fact that WFS-NMF considers the importance of different documents instead of treating them equally, the results of WFS-NMF achieves the best performance on most datasets.

B. Document Summarization

1) Data Sets: We use the DUC benchmark datasets (DUC2002 and DUC2004) for generic document summarization tasks. Table V gives a brief description of the data sets.

	DUC2002	DUC2004
number of document collections	59	50
number of documents in each collection	$\sim$10	10
data source	TREC	TDT
summary length	200 words	665bytes

Table V
DESCRIPTION OF THE DATA SETS FOR MULTI-DOCUMENT SUMMARIZATION

2) Implemented Systems: In this experiment, we compare our algorithms for summarization with several most widely used document summarization methods as follows.

- (1) ***DUCBest***: the method developed by the team achieving the highest scores in the DUC competition.
- (2)***Random***: selects sentences randomly for each document collection.
- (3) ***Centroid***: similar to MEAD algorithm proposed in [31] using centroid value, positional value, and first-sentence overlap as features.
- (4) ***LexPageRank***: a graph-based summarization method recommending sentences by the voting of their neighbors [13].
- (5) ***LSA***: conducts latent semantic analysis on terms by sentences matrix as proposed in [16].
- (6) ***NMF***: performs NMF on terms by sentences matrix and ranks the sentences by their weighted scores [24].

In order to use FS-NMF or WFS-NMF to conduct document summarization, we use the document-sentence matrix as the input data X, which can be generated from the document-term and sentence-term matrices, and now each feature (column) in X represents a sentence. Then the sentences can be ranked based on the sentence weights in W in both FS-NMF and WFS-NMF. Top-ranked sentences are included into the final summary. Since WFS-NMF weights both the samples and features, an alternative solution for document summarization is to factorize the sentence-term matrix generated from the original documents, and after computation the sentences are naturally ranked based on their assigned weights. Thus, we develop three new summarization methods as follows.

- (7) ***FS-NMF***: performs FS-NMF on document-sentence matrix, and selects the sentences associated with the highest weights to form summaries.
- (8) ***WFS-NMF-1***: similar to FS-NMF, performs WFS-NMF on document-sentence matrix to select the sentences with the highest weights.
- (9) ***WFS-NMF-2***: performs WFS-NMF on sentence-term matrix, and selects the sentences associated with the highest weights to form summaries.

3) Evaluation Methods: We use ROUGE [27] toolkit (version 1.5.5) to measure the summarization performance, which is widely applied by DUC for performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-W and ROUGE-SU. ROUGE-N is an n-gram recall computed as follows.

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{ref}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{ref}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \quad (23)$$

where n is the length of the n-gram, and ref stands for the reference summaries. $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of n-grams co-occurring in a candidate summary and the reference summaries, and $\text{Count}(\text{gram}_n)$ is the number of n-grams in the reference summaries. ROUGE-W is based on weighted LCS and ROUGE-SU is based on skip-bigram plus unigram. Each of these evaluation methods in ROUGE can generate three scores (recall, precision and F-measure). As we have similar conclusions in terms of any of the three scores, for simplicity, in this paper, we only report the average F-measure scores generated by ROUGE-1, ROUGE-2, ROUGE-W and ROUGE-SU to compare the implemented systems.

Systems	R-1	R-2	R-W	R-SU
DUC Best	**0.499**	0.252	0.201	0.284
Random	0.383	0.126	0.121	0.154
Centroid	0.454	0.192	0.180	0.180
LexPageRank	0.480	0.229	0.189	0.262
LSA	0.431	0.150	0.152	0.202
NMF	0.446	0.163	0.161	0.217
FS-NMF	0.493	0.249	0.201	**0.287**
WFS-NMF-1	**0.499**	**0.258**	**0.213**	**0.287**
WFS-NMF-2	0.491	0.252	0.198	0.283

Table VI
OVERALL PERFORMANCE COMPARISON ON DUC2002 DATA.

Systems	R-1	R-2	R-W	R-SU
DUC Best	0.382	0.092	0.133	0.132
Random	0.318	0.063	0.117	0.117
Centroid	0.367	0.361	0.124	0.125
LexPageRank	0.378	0.085	0.131	0.130
LSA	0.341	0.065	0.120	0.119
NMF	0.367	0.072	0.129	0.129
FS-NMF	0.388	0.101	0.139	0.134
WFS-NMF-1	**0.393**	**0.112**	**0.139**	**0.135**
WFS-NMF-2	0.388	0.103	0.137	**0.135**

Table VII
OVERALL PERFORMANCE COMPARISON ON DUC2004 DATA.

4) Summarization Evaluation: The experimental results are demonstrated in Table VI and Table VII. From the results, we have the following observations:

- All of the three summarization methods developed based on FS-NMF and WFS-NMF algorithms outperform the state-of-the-art generic summarization methods. The good results benefit from the weighting schemes for sentence features (or sentence samples).
- Among these three methods, in general WFS-NMF-1 achieves the highest ROUGE scores. This observation demonstrates that the sentence feature selection is effective and the weights on document side also helps the sentence weighting process.
- While further looking at the selected sentences, we find that there do exist some overlap in the selected sentences by the proposed three summarization methods, which indicates the consistency and effectiveness of the weight assignments in both samples and features.
- The ROUGE scores of our methods are higher than the best team in DUC2004 and comparable to the best team from DUC2002. Note that the good results of the best team come from the fact that they perform deeper natural language processing techniques to resolve pronouns and other anaphoric expressions, which we do not use for the data preprocessing.

C. Visualization

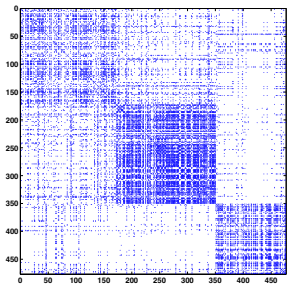
To evaluate the term features selected by our methods in document clustering and simultaneous keyword selection, in this set of experiments, we calculate the pairwise document similarity using the top 20 word features selected by different methods. We use CSTR dataset in this experiment, which contains four classes of text data. We compare the results of our FS-NMF and WFS-NMF algorithms with standard NMF and LSI, and Figure 1 demonstrates the document similarity matrix visually. Note that in the CSTR dataset, we order the documents based on their class labels.

From Figure 1, we have the following observations.

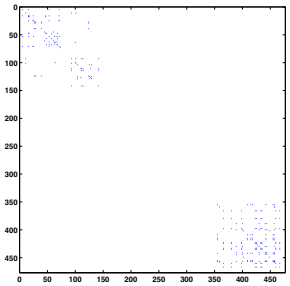
- Word features selected by FS-NMF and WFS-NMF can effectively reflet the document distribution. This is because the keywords identified by FS-NMF discriminate different topics in a global perspective.
- NMF Figure 1(b) shows no obvious patterns at all. The failure of NMF comes from the fact that it tries to group the terms into topics contained in the documents and uses the terms with the highest probabilities in each topic as the keywords, which are not discriminant and usually redundant.
- LSI can also find meaningful words, however, the first two clusters are not clearly discovered in Figure 1(a), which indicates some small classes are hard to identified by LSI using the keywords selected by it.

VIII. CONCLUSION

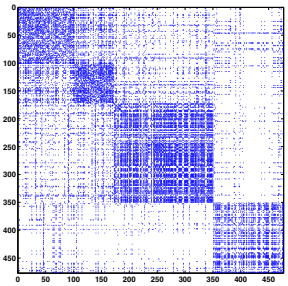
In this paper, we propose the weighed feature subset non-negative matrix factorization, which is an unsupervised approach to simultaneously cluster data points and select



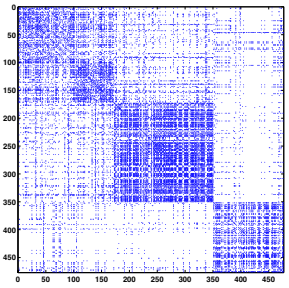
(a) LSI



(b) NMF



(c) WFS-NMF



(d) FS-NMF

Figure 1. Visualization Results on CSTR Data. CSTR has 4 clusters.

important features and also different data points are assigned different weights indicating their importance. We apply our proposed approach to various document understanding tasks including document clustering, summarization, and visualization. Experimental results demonstrate the effectiveness of our approaches for these tasks.

ACKNOWLEDGEMENT

The work of D. Wang is supported by an Florida International University (FIU) Dissertation Fellowship. The work of T. Li is partially supported NSF grants IIS-0546280, CCF-0830659, and DMS-0915110. The work of C. Ding is partially supported by NSF grants DMS-0844497 and CCF-0830780.

REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:9931022, 2003.

[2] D. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2:325–344, 1997.

[3] H. Cho, I. Dhillon, Y. Guan, and S. Sra. Minimum sum squared residue co-clustering of gene expression data. In *Proceedings of SDM 2004*.

[4] J. M. Conroy and D. P. O'leary. Text summarization via hidden markov models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407, 2001.

[5] I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of SIGKDD 2001*.

[6] I. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. 2004.

[7] I. Dhillon, S. Mallela, and S. Modha. Information-theoretic co-clustering. In *Proceedings of SIGKDD 2001*.

[8] C. Ding, X. He, and H. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of Siam Data Mining*, 2005.

[9] C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means c lustering. In *Proceedings of ICML 2007*.

[10] C. Ding, T. Li, and W. Peng. Nmf and plsi: equivalence and a hybrid algorithm. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 641–642, 2006.

[11] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of SIGKDD 2006*, 2006.

[12] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., 2001.

[13] G. Erkan and D. Radev. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP 2004*.

[14] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 601–602, 2005.

[15] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *In Research and Development in Information Retrieval*, pages 121–128, 1999.

[16] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of SIGIR 2001*.

[17] A. Graesser, A. Karnavat, and V. Pomeroy. Latent semantic analysis captures causal, goal-oriented, and taxonomic structures. In *CogSci*.

[18] Q. Gu and J. Zhou. Local learning regularized nonnegative matrix factorization. In *IJCAI*, pages 1046–1051, 2009.

[19] E.-H. S. Han, D. Bole, M. Gin, R. Gross, K. Hastings, G. Karypis, V. Kuma, B. Mobasher, and J. Moore. Webace: A web agent for document categorization and exploration, 1998.

[20] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[21] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[22] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.

[23] L. Jing, M. K. Ng, and J. Z. Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Trans. on Knowl. and Data Eng.*, 19(8):1026–1041, 2007.

[24] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562, 2001.

[25] T. Li. The relationships among various nonnegative matrix factorization methods for clustering. In *In ICDM*, pages 362–371, 2006.

[26] T. Li, S. Ma, and M. Ogihara. Document clustering via adaptive subspace iteration. In *Proceedings of Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 218–225, 2004.

[27] C.-Y. Lin and E.Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NLT-NAACL 2003*.

[28] C.-Y. Lin and E. Hovy. From single to multi-document summarization: a prototype system and its evaluation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 457–464, 2002.

[29] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[30] S. Park, J.-H. Lee, D.-H. Kim, and C.-M. Ahn. Multi-document summarization based on cluster using non-negative matrix factorization. In *SOFSEM '07: Proceedings of the 33rd conference on Current Trends in Theory and Practice of Computer Science*, pages 761–770, 2007.

[31] D. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, pages 919–938, 2004.

[32] E. Rennison. Galaxy of news: an approach to visualizing and understanding expansive news landscapes. In *UIST '94*, pages 3–12, 1994.

[33] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2862–2867, 2007.

[34] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 1997.

[35] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[36] A. Strehl, J. Ghosh, and C. Cardie. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[37] D. Wang, C. H. Q. Ding, and T. Li. Feature subset nonnegative matrix factorization and its applications to document understanding. In *SIGIR*, pages 805–806, 2010.

[38] D. Wang, T. Li, S. Zhu, and C. Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of SIGIR 2008*.

[39] F. Wang, C. Zhang, and T. Li. Regularized clustering for documents. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 95–102, 2007.

[40] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of SIGIR 2004*, 2003.

[41] W.-t. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. Multi-document summarization by maximizing informative content-words. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1776–1782, 2007.

[42] S. X. Yu and J. Shi. Multiclass spectral clustering. In *ICCV '03*.

[43] H. Zha, X. He, C. Ding, M. Gu, and H. Simon. Bipartite graph partitioning and data clustering. *Proc. Int'l Conf. Information and Knowledge Management (CIKM 2001)*, 2001.