Orthogonal Parametric Non-negative Matrix Tri-Factorization with α-Divergence for Co-clustering

Saeid Hoseinipour, Mina Aminghafari, Adel Mohammadpour

Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Iran.

saeidhoseinipour@aut.ac.ir (Saeid Hoseinipour)

aminghafari@aut.ac.ir (Mina Aminghafari)

adel@aut.ac.ir (Adel Mohammadpour)

Orthogonal Parametric Non-negative Matrix Tri-Factorization with α-Divergence for Co-clustering

Saeid Hoseinipour, Mina Aminghafari, Adel Mohammadpour

Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Iran.

Abstract

Co-clustering algorithms can seek homogeneous sub-matrices into a dyadic data matrix, such as a document-word matrix. Algorithms for co-clustering can be expressed as a non-negative matrix tri-factorization problem such that $\mathbf{X} \approx \mathbf{FSG}^\top$, which is associated with the non-negativity conditions on all matrices and the orthogonality of $\mathbf{F}$ (row-coefficient) and $\mathbf{G}$ (column-coefficient) matrices. Most algorithms are based on Euclidean distance and Kullback-Leibler divergence without parameters to control orthogonality. We propose to apply the orthogonality on parameters by adding two penalty terms based on the α-divergence objective function. Orthogonal parametric non-negative matrix tri-factorization uses orthogonal parameters for row and column space, separately. Finally, we compare the proposed algorithms with other algorithms on six real text datasets.

Keywords: Co-clustering, Dyadic data, α-divergence, Non-negative matrix tri-factorization, Orthogonality, Text mining.

2020 MSC: 15A23, 15B10

*Corresponding author

Email addresses: saeidhoseinipour@aut.ac.ir (Saeid Hoseinipour), aminghafari@aut.ac.ir (Mina Aminghafari), adel@aut.ac.ir (Adel Mohammadpour)

1. Introduction

Dyadic data matrices, such as document-word counts, movie-viewer ratings, and product-customer purchases matrices, refer to the duality between rows and columns that frequently arise in various essential applications—for example, collaborative filtering [1], text mining [2], and gene expression data analysis [3]. A fundamental problem in dyadic data analysis is discovering the hidden sub-matrices [4], [5]. Seeking sub-matrices in a dyadic data matrix is an old idea called co-clustering.

Traditional one-side clustering cannot seek a relation between sub-sets of rows (documents) and columns (words), but simultaneously clustering on rows and columns is useful for discovering sub-matrices into a dyadic data matrix [6], [7]. Matrix factorization methods have been widely used in dyadic data analysis [8], [9].

Non-negative Matrix Factorization (NMF) [10], [11], [12] and Non-negative Matrix Tri-Factorization (NMTF) are both important non-negative low rank approximation techniques. NMF is a framework for clustering, and co-clustering can be formulated by NMTF [13], [14].

Non-negative Block Value Decomposition (NBVD) [4] introduced a three non-negative factor decomposition for $\mathbf{X}$ as a document-word matrix by the multiplication of $\mathbf{F}$, $\mathbf{S}$, and $\mathbf{G}$ as the matrix of row coefficients, block value, and column coefficients, respectively. Orthogonal Non-negative Matrix 3-Factorization (ONM3F) is a three factorization of a non-negative $\mathbf{X}$ with orthogonality constraints ($\mathbf{F}^\top\mathbf{F} = \mathbf{I}_g$ and $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_s$) [8]. Orthogonal Non-negative Matrix Tri-Factorization (ONMTF) [9] developed a multiplicative update rule algorithm with orthogonality constraints. The co-clustering of document-word matrices by NBVD, ONM3F, and ONMTF was demon-

strated to be effective. Double NMTF (DNMTF) and Orthogonal Double NMTF (ODNMTF) were proposed in [15], with the idea of finding two factors matrices rather than three factors matrices, which summarized matrix $\mathbf{S} = \mathbf{F}^\top \mathbf{X}\mathbf{G}$. Penalized Non-negative Matrix Tri-Factorization (PNMTF) introduced three penalty terms to replace the three orthogonality constraints [16].

These methods find factor matrices $\mathbf{F}$ and $\mathbf{G}$ based on Euclidean distance or Kullback–Leibler divergence as an objective function. Orthogonality of $\mathbf{F}$ and $\mathbf{G}$ guarantees the uniqueness of the matrices [8] and also applies independent constraint among labels of rows (columns) from a statistical perspective.

In recent years, various measures of dissimilarity, such as Bregman divergences, Csiszár's f-divergences, and Amari's α-divergences, have been examined under NMF [17], [18], [19], [20]. Multiplicative update rules algorithms were proposed in [21], based on α-divergence [22] as a particular case of Csiszár's f-divergence. NMF multiplicative algorithm based on α-divergence was proposed in [17] for image denoising and EEG classification. Suppose that $\mathbf{X}$ is a document-word matrix, which factor matrices obtained using (1) are represented in Figure 1.

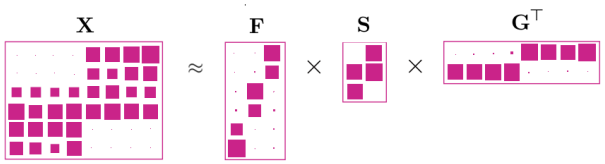

Figure 1: Similar to [9], a graphical illustration of document-word co-clustering $\mathbf{X} \in \mathbb{R}_+^{6\times 8}$. An element in the matrix with a bigger square has a larger value. NMTF identifies three factors $\mathbf{F}$, $\mathbf{S}$, and $\mathbf{G}$ for three rows (document) clusters and two columns (word) clusters. $\mathbf{S}$ is a matrix that shows the connections between document-clusters and word-clusters. Each column of $\mathbf{S}$ indicates which document-clusters contribute to the word-cluster. The second and third document-cluster are related to first word-cluster, also the first and second document-cluster are related to second word-cluster.

In this paper, we present NMTF based on α-divergence in Section 2. We introduce a novel orthogonal parametric non-negative matrix tri-factorization in Section 3, with convergence proof and Multiplicative update rules by matrix-wise form. Also, we demonstrate that our algorithm is derivable using the Karush-Kuhn-Tucker conditions. In Section 4, we implement competitor algorithms listed in Table 1 on six real text datasets. All notations are listed in Appendix A.

2. Non-negative Matrix Tri-Factorization

Given observation matrix $\mathbf{X} = (X_{ij}) \in \mathbb{R}_+^{n\times m}$ and approximation matrix $\mathbf{FSG}^\top = ([\mathbf{FSG}^\top]_{ij}) \in \mathbb{R}_+^{n\times m}$. NMTF aims to find three matrices $\mathbf{F} = (F_{ik}) \in \mathbb{R}_+^{n\times g}$, $\mathbf{S} = (S_{kh}) \in \mathbb{R}_+^{g\times s}$, and $\mathbf{G} = (G_{jh}) \in \mathbb{R}_+^{m\times s}$ with non-negative elements. The low-rank approximation of $\mathbf{X}$ by

$$\mathbf{X} \approx \mathbf{FSG}^\top, \tag{1}$$

where n, m, $g \leqslant n$, and $s \leqslant m$ are the numbers of rows, columns, row clusters, and column clusters, respectively. Also, $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{S}$ play roles

Table 1: The comparison between algorithms according to the optimization problem and multiplicative update rules.

Algorithm	Optimization problem	Multiplicative update rules
OPNMTF (Proposed 3)	$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0}{\arg\min} D_\alpha(\mathbf{X}\|\mathbf{FSG}^\top)+\lambda D_\alpha(\mathbf{I}_g\|\mathbf{F}^\top\mathbf{F})+\mu D_\alpha(\mathbf{I}_s\|\mathbf{G}^\top\mathbf{G})$	$\mathbf{F}\odot\left[\left(\frac{(\mathbf{X}\oslash\mathbf{FSG}^\top)^\alpha\mathbf{GS}^\top+2\lambda\mathbf{F}(\mathbf{I}_g\oslash\mathbf{F}^\top\mathbf{F})^\alpha}{\mathbf{E}_{mn}\mathbf{GS}^\top+2\lambda\mathbf{FI}_g}\right)^{\frac{1}{\alpha}}\right]_+$, $\mathbf{G}\odot\left[\left(\frac{((\mathbf{X}\oslash\mathbf{FSG}^\top)^\top)^\alpha\mathbf{FS}+2\mu\mathbf{G}(\mathbf{I}_s\oslash\mathbf{G}^\top\mathbf{G})^\alpha}{\mathbf{E}_{nn}^\top\mathbf{FS}+2\mu\mathbf{GI}_s}\right)^{\frac{1}{\alpha}}\right]_+$, $\mathbf{S}\odot\left[\left(\frac{\mathbf{F}^\top(\mathbf{X}\oslash\mathbf{FSG}^\top)^\alpha\mathbf{G}}{\mathbf{F}^\top\mathbf{E}_{nn}\mathbf{G}}\right)^{\frac{1}{\alpha}}\right]_+$.
PNMTF [16]	$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0}{\arg\min}\frac{1}{2}\|\mathbf{X}-\mathbf{FSG}^\top\|^2+\frac{\tau}{2}\mathrm{Tr}(\mathbf{F}\mathbf{\Psi}_g\mathbf{F}^\top)+\frac{\eta}{2}\mathrm{Tr}(\mathbf{G}\mathbf{\Psi}_s\mathbf{G}^\top)+\frac{\gamma}{2}\mathrm{Tr}(\mathbf{S}^\top\mathbf{S})$	$\mathbf{F}\odot\left[\left(\frac{\mathbf{XGS}^\top}{\mathbf{FSG}^\top\mathbf{GS}^\top+\tau\mathbf{F}\mathbf{\Psi}_g}\right)^{\frac{1}{2}}\right]_+$, $\mathbf{G}\odot\left[\left(\frac{\mathbf{X}^\top\mathbf{FS}}{\mathbf{GS}^\top\mathbf{F}^\top\mathbf{FS}+\eta\mathbf{G}\mathbf{\Psi}_s}\right)^{\frac{1}{2}}\right]_+$, $\mathbf{S}\odot\left[\left(\frac{\mathbf{F}^\top\mathbf{XG}}{\mathbf{F}^\top\mathbf{FSG}^\top\mathbf{G}+\gamma\mathbf{S}}\right)^{\frac{1}{2}}\right]_+$.
DNMF [15]	$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0,\ \mathbf{F}^\top\mathbf{F}=\mathbf{I}_g,\mathbf{G}^\top\mathbf{G}=\mathbf{I}_s}{\arg\min}\|\mathbf{X}-\mathbf{FF}^\top\mathbf{XGG}^\top\|^2$	$\mathbf{F}\odot\left[\frac{\mathbf{XGG}^\top\mathbf{X}^\top\mathbf{F}}{\mathbf{FF}^\top\mathbf{XGG}^\top\mathbf{X}^\top\mathbf{F}}\right]_+$, $\mathbf{G}\odot\left[\frac{\mathbf{X}^\top\mathbf{FF}^\top\mathbf{XG}}{\mathbf{GG}^\top\mathbf{X}^\top\mathbf{FF}^\top\mathbf{XG}}\right]_+$.
ODNMF [15]	$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0}{\arg\min}\|\mathbf{X}-\mathbf{FF}^\top\mathbf{XGG}^\top\|^2+\mathrm{Tr}(\mathbf{\Lambda F}^\top)+\mathrm{Tr}(\mathbf{\Gamma G}^\top)$	$\mathbf{F}\odot\left[\frac{2\mathbf{XGG}^\top\mathbf{X}^\top\mathbf{F}}{\mathbf{FF}^\top\mathbf{XGG}^\top\mathbf{GG}^\top\mathbf{X}^\top\mathbf{F}+\mathbf{XGG}^\top\mathbf{GG}^\top\mathbf{X}^\top\mathbf{FF}^\top\mathbf{F}}\right]_+$, $\mathbf{G}\odot\left[\frac{2\mathbf{X}^\top\mathbf{FF}^\top\mathbf{XG}}{\mathbf{GG}^\top\mathbf{FF}^\top\mathbf{XX}^\top\mathbf{FF}^\top\mathbf{G}+\mathbf{FF}^\top\mathbf{XX}^\top\mathbf{FF}^\top\mathbf{GG}^\top\mathbf{G}}\right]_+$.
ONMTF [9]	$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0,\ \mathbf{F}^\top\mathbf{F}=\mathbf{I}_g,\mathbf{G}^\top\mathbf{G}=\mathbf{I}_s}{\arg\min}\frac{1}{2}\|\mathbf{X}-\mathbf{FSG}^\top\|^2$	$\mathbf{F}\odot\left[\frac{\mathbf{XGS}^\top}{\mathbf{FSG}^\top\mathbf{X}^\top\mathbf{F}}\right]_+$, $\mathbf{G}\odot\left[\frac{\mathbf{X}^\top\mathbf{FS}}{\mathbf{GS}^\top\mathbf{F}^\top\mathbf{XG}}\right]_+$, $\mathbf{S}\odot\left[\frac{\mathbf{F}^\top\mathbf{XG}}{\mathbf{F}^\top\mathbf{FSG}^\top\mathbf{G}}\right]_+$.
ONM3F [8]	$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0,\ \mathbf{F}^\top\mathbf{F}=\mathbf{I}_g,\mathbf{G}^\top\mathbf{G}=\mathbf{I}_s}{\arg\min}\|\mathbf{X}-\mathbf{FSG}^\top\|^2+\mathrm{Tr}(\mathbf{\Lambda}(\mathbf{F}^\top\mathbf{F}-\mathbf{I}_s))+\mathrm{Tr}(\mathbf{\Gamma}(\mathbf{G}^\top\mathbf{G}-\mathbf{I}_g))$	$\mathbf{F}\odot\left[\left(\frac{\mathbf{XGS}^\top}{\mathbf{FF}^\top\mathbf{XGS}^\top}\right)^{\frac{1}{2}}\right]_+$, $\mathbf{G}\odot\left[\left(\frac{\mathbf{X}^\top\mathbf{FS}}{\mathbf{GG}^\top\mathbf{X}^\top\mathbf{FS}}\right)^{\frac{1}{2}}\right]_+$, $\mathbf{S}\odot\left[\left(\frac{\mathbf{F}^\top\mathbf{XG}}{\mathbf{F}^\top\mathbf{FSG}^\top\mathbf{G}}\right)^{\frac{1}{2}}\right]_+$.
NBVD [4]	$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0}{\arg\min}\|\mathbf{X}-\mathbf{FSG}^\top\|^2$	$\mathbf{F}\odot\left[\frac{\mathbf{XGS}^\top}{\mathbf{FSG}^\top\mathbf{GS}^\top}\right]_+$, $\mathbf{G}\odot\left[\frac{\mathbf{X}^\top\mathbf{FS}}{\mathbf{GS}^\top\mathbf{F}^\top\mathbf{FS}}\right]_+$, $\mathbf{S}\odot\left[\frac{\mathbf{F}^\top\mathbf{XG}}{\mathbf{F}^\top\mathbf{FSG}^\top\mathbf{G}}\right]_+$.

* [16] τ, η, and γ are parameters orthogonality, also $\mathbf{\Psi}_g=\mathbf{E}_{gg}-\mathbf{I}_g$ and $\mathbf{\Psi}_s=\mathbf{E}_{ss}-\mathbf{I}_s$. [15] $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ are matrices containing Lagrangian multipliers. [4] proposed two versions when $\mathbf{S}=\mathbf{F}^\top\mathbf{XG}$, also DNMF is called Double k-means.

in membership row clustered, column clustered, and summarization matrix, respectively. This triple decomposition enables a suitable framework for simultaneous clustering on $\mathbf{X}$.

Definition 1. The α-divergence between $\mathbf{X}$ and $\mathbf{FSG}^\top$ is defined as follows [17]:

$$
\begin{aligned}
D_\alpha(\mathbf{X}\|\mathbf{FSG}^\top) &= \frac{1}{\alpha(1-\alpha)}\sum_{i,j}\left(\alpha X_{ij}+(1-\alpha)[\mathbf{FSG}^\top]_{ij}-X_{ij}^\alpha[\mathbf{FSG}^\top]_{ij}^{1-\alpha}\right)\\
&= \sum_{i,j}X_{ij}f\left(\frac{[\mathbf{FSG}^\top]_{ij}}{X_{ij}}\right), \qquad (2)
\end{aligned}
$$

where $f(.)$ is a convex function for positive values α as follows

$$f(y) = \frac{1}{\alpha(1-\alpha)} \left(\alpha + (1-\alpha)y - y^{1-\alpha}\right). \tag{3}$$

The α-divergence (2) includes Kullback-Leibler (KL), Hellinger, and Person-χ^2 divergences. The divergence criteria for special values of α are listed in Table 2, and their corresponding functions (3) are plotted in Figure 2.

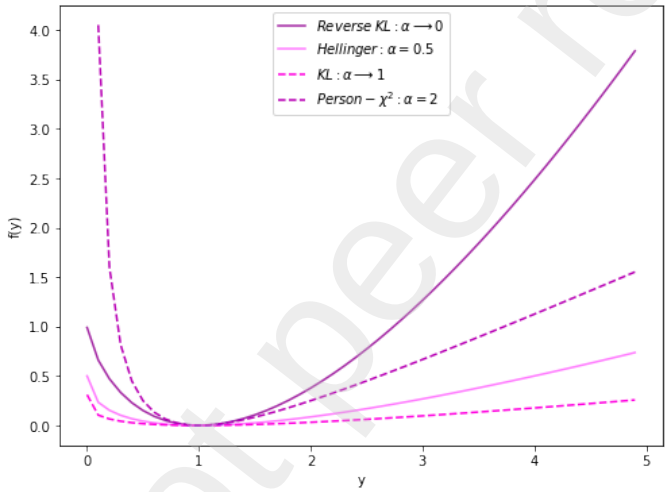

Figure 2: The convex function $f(y)$ in (3) for various values.

Table 2: Types of α-divergence for $\mathbf{X}$ as data matrix and $\mathbf{FSG}^\top$ as approximation matrix.

Divergence	α	$D_\alpha(\mathbf{X}\|\mathbf{FSG}^\top)$
KL	1^*	$\sum_{i,j}\left(X_{ij}\log(\frac{X_{ij}}{[\mathbf{FSG}^\top]_{ij}}) - X_{ij} + [\mathbf{FSG}^\top]_{ij}\right)$
Hellinger	0.5	$2\sum_{i,j}\left(\sqrt{X_{ij}} - \sqrt{[\mathbf{FSG}^\top]_{ij}}\right)^2$
Person-χ^2	2	$\frac{1}{2}\sum_{i,j}\frac{\left(X_{ij} - [\mathbf{FSG}^\top]_{ij}\right)^2}{[\mathbf{FSG}^\top]_{ij}}$

* $\lim_{\alpha\to 1} D_\alpha(\mathbf{X}||\mathbf{FSG}^\top)$ equal to KL, and $\lim_{\alpha\to 0} D_\alpha(\mathbf{X}||\mathbf{FSG}^\top)$ equal to reverse KL.

Multiplicative update rules are efficient algorithms for solving the underlying optimization problem

$$\underset{\mathbf{F}\geqslant 0,\mathbf{S}\geqslant 0,\mathbf{G}\geqslant 0}{\arg\min} \; D_\alpha(\mathbf{X}\|\mathbf{F}\mathbf{S}\mathbf{G}^\top). \tag{4}$$

It is crucial to introduce an auxiliary function to analyze convergence and to derive algorithms.

Definition 2. $A(\mathbf{F}, \mathbf{F}^{(t)})$ is called an auxiliary function for $D_\alpha(\mathbf{X}\|\mathbf{F}\mathbf{S}\mathbf{G}^\top)$ as a function of $\mathbf{F}$ if conditions (5) and (6) are satisfied:

$$A(\mathbf{F}, \mathbf{F}) = D_\alpha(\mathbf{X}\|\mathbf{F}\mathbf{S}\mathbf{G}^\top), \tag{5}$$

$$A(\mathbf{F}, \mathbf{F}^{(t)}) \geq D_\alpha(\mathbf{X}\|\mathbf{F}\mathbf{S}\mathbf{G}^\top), \tag{6}$$

where $\mathbf{F}^{(t)}$ denotes the iteration t that resulted in the updating of $\mathbf{F}$.

Lemma 2.1. *If $A(\mathbf{F}^{(t+1)}, \mathbf{F}^{(t)})$ is an auxiliary function for $D_\alpha(\mathbf{X}\|\mathbf{F}^{(t+1)}\mathbf{S}\mathbf{G}^\top)$, then $D_\alpha(\mathbf{X}\|\mathbf{F}^{(t+1)}\mathbf{S}\mathbf{G}^\top)$ is non-increasing function of $\mathbf{F}^{(t+1)}$ with respect to the updating rule:*

$$\mathbf{F}^{(t+1)} = \arg\min_{\mathbf{F}} A(\mathbf{F}, \mathbf{F}^{(t)}). \tag{7}$$

Proof. See Appendix B. □

3. Orthogonal Parametric Non-negative Matrix Tri-Factorization

Orthogonality $\mathbf{F}$ and $\mathbf{G}$ can be achieved without considering restrictions ($\mathbf{F}^\top\mathbf{F} = \mathbf{I}_g$, $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_s$). This goal is applied in Orthogonal Parametric Non-negative Matrix Tri-Factorization (OPNMTF). The objective function in OPNMTF has a two-term penalty based on α-divergence. The first term is for the row space with the control parameter λ. The second term is for

the column space with the control parameter μ. We introduce our objective function:

$$\xi_{\text{OPNMTF}} = D_\alpha(\mathbf{X}\|\mathbf{FSG}^\top) + \lambda D_\alpha(\mathbf{I}_g\|\mathbf{F}^\top\mathbf{F}) + \mu D_\alpha(\mathbf{I}_s\|\mathbf{G}^\top\mathbf{G}) \tag{8}$$

Lemma 3.1. *If* $\mathbf{G}$ *and* $\mathbf{S}$ *are fixed, then the auxiliary function for* ξ_{OPNMTF} *is*

$$A(\mathbf{F}, \mathbf{F}^{(t)}) = \sum_{i,j,k,h} X_{ij} Q^{\mathbf{F}^{(t)}}_{ijkh}\, f\left(\frac{F_{ik}S_{kh}G^\top_{hj}}{X_{ij}Q^{\mathbf{F}^{(t)}}_{ijkh}}\right) + \lambda \sum_{i,k} Q^{\mathbf{F}^{(t)}}_{ik}\, f\left(\frac{F^\top_{ki}F_{ik}}{Q^{\mathbf{F}^{(t)}}_{ik}}\right) + Const_{\mathbf{G}}, \tag{9}$$

where $Q^{\mathbf{F}^{(t)}}_{ijkh} = \dfrac{F^{(t)}_{ik}S_{kh}G^\top_{hj}}{\sum_{k',h'} F^{(t)}_{ik'}S_{k'h'}G^\top_{h'j}}$, $Q^{\mathbf{F}^{(t)}}_{ik} = \dfrac{F^{(t)\top}_{ki}F^{(t)}_{ik}}{\sum_{i'} F^{(t)\top}_{ki'}F^{(t)}_{i'k}}$, *and* $Const_{\mathbf{G}}$ *is a constant value and depends on* $\mathbf{G}$, *and* $f(.)$ *is defined in* (3).

Proof. See Appendix C. □

Lemma 3.2. *If* $\mathbf{F}$ *and* $\mathbf{S}$ *are fixed, then the auxiliary function for* ξ_{OPNMTF} *is*

$$A(\mathbf{G}, \mathbf{G}^{(t)}) = \sum_{i,j,k,h} X_{ij} Q^{\mathbf{G}^{(t)}}_{ijkh}\, f\left(\frac{F_{ik}S_{kh}G^\top_{hj}}{X_{ij}Q^{\mathbf{G}^{(t)}}_{ijkh}}\right) + \mu \sum_{j,h} Q^{\mathbf{G}^{(t)}}_{jh}\, f\left(\frac{G^\top_{hj}G_{jh}}{Q^{\mathbf{G}^{(t)}}_{jh}}\right) + Const_{\mathbf{F}}, \tag{10}$$

where $Q^{\mathbf{G}^{(t)}}_{ijkh} = \dfrac{F_{ik}S_{kh}G^{(t)\top}_{hj}}{\sum_{k',h'} F_{ik'}S_{k'h'}G^{(t)\top}_{h'j}}$, $Q^{\mathbf{G}^{(t)}}_{jh} = \dfrac{G^{(t)\top}_{hj}G^{(t)}_{jh}}{\sum_{j'} G^{(t)\top}_{hj'}G^{(t)}_{j'h}}$, *and* $Const_{\mathbf{F}}$ *is a constant value and depends on* $\mathbf{F}$, *and* $f(.)$ *is defined in* (3).

Proof. It is the same as Lemma 3.1. □

Lemma 3.3. *If* $\mathbf{F}$ *and* $\mathbf{G}$ *are fixed, then the auxiliary function for* ξ_{OPNMTF} *is*

$$A(\mathbf{S}, \mathbf{S}^{(t)}) = \sum_{i,j,k,h} X_{ij} Q^{\mathbf{S}^{(t)}}_{ijkh}\, f\left(\frac{F_{ik}S_{kh}G^\top_{hj}}{X_{ij}Q^{\mathbf{S}^{(t)}}_{ijkh}}\right) + Const_{\mathbf{F},\mathbf{G}}, \tag{11}$$

where $Q_{ijkh}^{\mathbf{S}^{(t)}} = \dfrac{F_{ik}S_{kh}^{(t)}G_{hj}^{\top}}{\sum\limits_{k',h'} F_{ik'}S_{k'h'}^{(t)}G_{h'j}^{\top}}$, $Const_{\mathbf{F},\mathbf{G}}$ *is a constant value and depends on* $\mathbf{F}$ *and* $\mathbf{G}$, *and* $f(.)$ *is defined in* (3).

Proof. It is the same as Lemma 3.1. □

Theorem 3.4. *The multiplicative update rule for* ξ_{OPNMTF} (8) *is given as follows:*

$$\mathbf{F}^{(t+1)} \leftarrow \mathbf{F}^{(t)} \odot \left[\left(\frac{(\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\alpha}\mathbf{G}\mathbf{S}^{\top} + 2\lambda\mathbf{F}(\mathbf{I}_g \oslash \mathbf{F}^{\top}\mathbf{F})^{\alpha}}{\mathbf{E}_{nm}\mathbf{G}\mathbf{S}^{\top} + 2\lambda\mathbf{F}\mathbf{I}_g}\right)^{\frac{1}{\alpha}}\right]_{+} \tag{12}$$

$$\mathbf{G}^{(t+1)} \leftarrow \mathbf{G}^{(t)} \odot \left[\left(\frac{((\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\top})^{\alpha}\mathbf{F}\mathbf{S} + 2\mu\mathbf{G}(\mathbf{I}_s \oslash \mathbf{G}^{\top}\mathbf{G})^{\alpha}}{\mathbf{E}_{nm}^{\top}\mathbf{F}\mathbf{S} + 2\mu\mathbf{G}\mathbf{I}_s}\right)^{\frac{1}{\alpha}}\right]_{+} \tag{13}$$

$$\mathbf{S}^{(t+1)} \leftarrow \mathbf{S}^{(t)} \odot \left[\left(\frac{\mathbf{F}^{\top}(\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\alpha}\mathbf{G}}{\mathbf{F}^{\top}\mathbf{E}_{nm}\mathbf{G}}\right)^{\frac{1}{\alpha}}\right]_{+} \tag{14}$$

where $[\mathbf{P}]_{+} = \max(\mathbf{P}, 0)$, $\odot$ *and* $\oslash$ *are element-wise multiplication and division, respectively.* $\mathbf{E}_{nm}$ *is a matrix of ones with sizes* $n \times m$. *Also,* $\mathbf{I}_g$ *and* $\mathbf{I}_s$ *are identity matrices with sizes* $g \times g$ *and* $s \times s$, *respectively. The parameters* $\lambda \geqslant 0$ *and* $\mu \geqslant 0$ *adjust the orthogonality of the vectors in* $\mathbf{F}$ *and* $\mathbf{G}$.

Proof. See Appendix D. □

3.1. Karush–Kuhn–Tucker conditions

The Karush–Kuhn–Tucker (KKT) conditions [23] define a first-order constraint that must be met when we solve non-linear optimization problems. The minimization of ξ_{OPNMTF} in (8) can be defined as an inequality constrained minimization problem. It is possible to determine (12), (13), and (14) using the KKT conditions.

Algorithm 1 Orthogonal Parametric Non-negative Matrix Tri-Factorization with α-divergence (OPNMTF)

Input: $\mathbf{X} \in \mathbb{R}_+^{n \times m}$ (data matrix),
g (number of row clusters),
s (number of column clusters),
α (divergence parameter),
λ (row orthogonality parameters),
μ (column orthogonality parameters),
N (maximum number of iterations).

Initialize: $\mathbf{F}^{(0)}$ and $\mathbf{G}^{(0)}$. Compute $\mathbf{S}^{(0)} = {\mathbf{F}^{(0)}}^\top \mathbf{X}\mathbf{G}^{(0)}$.

While not convergent and $1 \leqslant t \leqslant N$ do:
    Update $\mathbf{F}^{(t)}$ from (12),
    Update $\mathbf{S}^{(t)}$ from (13),
    Update $\mathbf{G}^{(t)}$ from (14),
    Set $t = t + 1$.
Normalize $\mathbf{F}$, $\mathbf{S}$, and $\mathbf{G}$ with probabilistic interpretation [9]:
    $\mathbf{F} \longleftarrow \mathbf{F}\mathbf{D}_\mathbf{F}^{-1}$,
    $\mathbf{S} \longleftarrow \mathbf{D}_\mathbf{F}\mathbf{S}\mathbf{D}_\mathbf{G}^{-1}$,
    $\mathbf{G} \longleftarrow \mathbf{G}\mathbf{D}_\mathbf{G}$,
where $\mathbf{D}_\mathbf{F} = \mathrm{diag}(\mathbf{1}_n^\top \mathbf{F})$ and $\mathbf{D}_\mathbf{G} = \mathrm{diag}(\mathbf{1}_m^\top \mathbf{G})$.
Assign row i and column j to row and column cluster k^* and h^* if
    $k^* = \arg\max_k F_{ik}, \qquad h^* = \arg\max_h G_{jh}$.

Output: $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{S}$.

Optimality conditions in KKT are

$$\frac{\partial \xi_{\mathrm{OPNMTF}}}{\partial \mathbf{F}} \odot \mathbf{F} = \mathbf{0},$$
$$\frac{\partial \xi_{\mathrm{OPNMTF}}}{\partial \mathbf{G}} \odot \mathbf{G} = \mathbf{0},$$
$$\frac{\partial \xi_{\mathrm{OPNMTF}}}{\partial \mathbf{S}} \odot \mathbf{S} = \mathbf{0},$$

and the conditions of complementary slackness are suggested by

$$\begin{aligned}
\frac{\partial \xi_{\text{OPNMTF}}}{\partial \mathbf{F}} \odot \mathbf{F}^{\alpha} &= \mathbf{0}, \\
\frac{\partial \xi_{\text{OPNMTF}}}{\partial \mathbf{G}} \odot \mathbf{G}^{\alpha} &= \mathbf{0}, \\
\frac{\partial \xi_{\text{OPNMTF}}}{\partial \mathbf{S}} \odot \mathbf{S}^{\alpha} &= \mathbf{0}.
\end{aligned}$$

The derivatives of ξ_{OPNMTF} over $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{S}$ are

$$\begin{aligned}
\frac{\partial \xi_{\text{OPNMTF}}}{\partial \mathbf{F}} &= \frac{1}{\alpha}\left(\mathbf{E}_{nm}\mathbf{G}\mathbf{S}^{\top} + 2\lambda\mathbf{F}\mathbf{I}_g\right) - \frac{1}{\alpha}\left((\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\alpha}\mathbf{G}\mathbf{S}^{\top} + 2\lambda\mathbf{F}(\mathbf{I}_g \oslash \mathbf{F}^{\top}\mathbf{F})^{\alpha}\right), \\
\frac{\partial \xi_{\text{OPNMTF}}}{\partial \mathbf{G}} &= \frac{1}{\alpha}\left(\mathbf{E}_{nm}^{\top}\mathbf{F}\mathbf{S} + 2\mu\mathbf{G}\mathbf{I}_s\right) - \frac{1}{\alpha}\left(\left((\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\top}\right)^{\alpha}\mathbf{F}\mathbf{S} + 2\mu\mathbf{G}(\mathbf{I}_s \oslash \mathbf{G}^{\top}\mathbf{G})^{\alpha}\right), \\
\frac{\partial \xi_{\text{OPNMTF}}}{\partial \mathbf{S}} &= \mathbf{F}^{\top}\mathbf{E}_{nm}\mathbf{G} - \mathbf{F}^{\top}(\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\alpha}\mathbf{G}.
\end{aligned}$$

The KKT condition leads us to the following equations:

$$\begin{aligned}
&\frac{1}{\alpha}\left(\mathbf{E}_{nm}\mathbf{G}\mathbf{S}^{\top} + 2\lambda\mathbf{F}\mathbf{I}_g\right)\mathbf{F}^{\alpha} - \frac{1}{\alpha}\left((\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\alpha}\mathbf{G}\mathbf{S}^{\top} + 2\lambda\mathbf{F}(\mathbf{I}_g \oslash \mathbf{F}^{\top}\mathbf{F})^{\alpha}\right)\mathbf{F}^{\alpha} = \mathbf{0}, \\
&\frac{1}{\alpha}\left(\mathbf{E}_{nm}^{\top}\mathbf{F}\mathbf{S} + 2\mu\mathbf{G}\mathbf{I}_s\right)\mathbf{G}^{\alpha} - \frac{1}{\alpha}\left(\left((\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\top}\right)^{\alpha}\mathbf{F}\mathbf{S} + 2\mu\mathbf{G}(\mathbf{I}_s \oslash \mathbf{G}^{\top}\mathbf{G})^{\alpha}\right)\mathbf{G}^{\alpha} = \mathbf{0}, \\
&\left(\mathbf{F}^{\top}\mathbf{E}_{nm}\mathbf{G} - \mathbf{F}^{\top}(\mathbf{X} \oslash \mathbf{F}\mathbf{S}\mathbf{G}^{\top})^{\alpha}\mathbf{G}\right)\mathbf{S}^{\alpha} = \mathbf{0}.
\end{aligned}$$

Therefore, the multiplicative update rule for (8) provided by equations (12), (13), and (14).

4. Numerical Experiments

In real text datasets, due to not having true column labels, we evaluate the quality of row clustering by one internal and three external measures. The algorithms performances in Table 1 are compared by three following criteria: Accuracy (Acc), Normalize Mutual Information (NMI), and Adjusted Rand

Index (ARI). Also, Inter-cluster Centroids Average Similarity (ICAS) as an internal measure can be used for datasets without row and column true labels.

4.1. Evaluation Measures

Acc [24] is the basic criteria for supervised clustering and it is defined as:

$$\mathrm{Acc} = \frac{1}{n} \max \left[\sum_{\mathcal{C}_k, \mathcal{L}_{k'}} \mathrm{T}(\mathcal{C}_k, \mathcal{L}_{k'}) \right],$$

where $\mathcal{C}_k$, $\mathcal{L}_{k'}$, and $\mathrm{T}(\mathcal{C}_k, \mathcal{L}_{k'}) = \mathcal{C}_k \cap \mathcal{L}_{k'}$ are sets that defines

$\mathcal{C}_k$	the prediction k-th cluster,
$\mathcal{L}_{k'}$	the true k'-th class,
$\mathrm{T}(\mathcal{C}_k, \mathcal{L}_{k'})$	the proportion of correctly recovered objects.

NMI [25] is calculated as follows:

$$\mathrm{NMI} = \frac{\sum\limits_{k,k'} \frac{n_{kk'}}{n} \log \frac{n_{kk'}}{n_k \hat{n}_{k'}}}{\sqrt{\left(\sum\limits_{k} \frac{n_k}{n} \log \frac{n_k}{n} \right) \left(\sum\limits_{k'} \frac{\hat{n}_{k'}}{n} \log \frac{\hat{n}_{k'}}{n} \right)}},$$

where n, n_k, $\hat{n}_{k'}$, and $n_{kk'}$ are integer numbers define as follows

n	# rows of data matrix $\mathbf{X}$,
n_k	# data in the cluster $\mathcal{C}_k (1 \leqslant k \leqslant g)$,
$\hat{n}_{k'}$	# data in the class $\mathcal{L}_{k'} (1 \leqslant k' \leqslant g)$,
$n_{kk'}$	# data in $\mathcal{C}_k \cap \mathcal{L}_{k'}$.

ARI [26] computes the similarity between two clustering partitions de-

fined as follows:

$$\text{ARI} = \frac{\sum_{k,k'} \binom{n_{kk'}}{2} - \left[\sum_{k} \binom{n_k}{2} \sum_{h} \binom{\hat{n}_{k'}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_{k} \binom{n_k}{2} + \sum_{k'} \binom{\hat{n}_{k'}}{2}\right] - \left[\sum_{k} \binom{n_k}{2} \sum_{k'} \binom{\hat{n}_{k'}}{2}\right] / \binom{n}{2}}.$$

The ARI measures the similarity between estimated and true cluster labels. These three metrics have a range between 0 and 1, and a value close to 1 indicates a good clustering result.

ICAS [5] is used to measure similarity between cluster centroids (i.e., the columns of $\mathbf{F}$) as follows:

$$\text{ICAS} = \frac{\sum_{k=1}^{g} \sum_{k'=k+1}^{g} \text{sim}\left(\text{centroid}_k, \text{centroid}_{k'}\right)}{0.5 \times g \times (g-1)}, \tag{15}$$

where the centroid_k is the k-th cluster centroid, $\text{sim}(\text{centroid}_k, \text{centroid}_{k'})$ is the cosine similarity function, and g is the number of row-clusters. This measure is between 0 and 1, and the optimal value is the minimum.

4.2. Real Datasets

The proposed algorithms were compared on six text datasets, which are presented in Table 3.

Table 3: Description of datasets by topics, number of documents and words.

Datasets	**Topics**	(#Documents, #Words)
CSTR [27]	Natural Language Processing, Robotics/Vision, Systems, and Theory	(475,1000)
WebACE [28]	20 different topics were obtained from the WebACE project	(2340,1000)
RCV1 [29]	It is a sub-set of a newswire stories corpus made available by Reuters containing four categories: C15, ECAT, GCAT, and MCAT.	(9625,29992)
Reviews [30]	Food, Music, Movies, Radio and Restaurants.	(4069,18483)
Sports [30]	Baseball, Boxing, Basketball, Bicycling, Football, Golfing, and Hockey.	(8580,14870)
Classic3 [31]	Medical, Information retrieval, and Aeronautical systems.	(3891,4303)

4.3. Results

OPNMTF is the best algorithm for clustering performance in **CSTR** and **Classic3**. Algorithms in Table 1 are compared with OPNMTF in Table 4. Figure 3 displays the reorganization matrices of clustering and co-clustering by implementing the OPNMTF algorithm for **Classic3**.

Table 4: Performance comparison of all algorithms on six real text datasets by averaging over 100 iterations.

Datasets	Metric	Algorithms						
		NBVD	ONM3F	ONMTF	ODNMTF	DNMTF	PNMTF	OPNMTF
CSTR	Acc	0.86	0.84	0.90	0.78	0.85	0.78	**0.93**
	NMI	0.78	0.76	0.78	0.73	0.80	0.69	**0.85**
	ARI	0.73	0.68	0.74	0.64	0.72	0.59	**0.83**
	ICAS	**0.29**	0.32	0.35	0.34	0.34	0.78	0.30
WebACE	Acc	0.29	0.30	0.27	0.37	**0.39**	0.31	0.33
	NMI	0.76	0.76	0.72	0.76	0.77	0.78	**0.79**
	ARI	0.43	0.44	0.40	0.44	0.46	0.48	**0.49**
	ICAS	0.30	0.31	**0.25**	0.37	0.35	0.80	0.27
RCV1	Acc	0.90	0.84	0.86	0.86	0.86	**0.91**	0.90
	NMI	0.80	0.73	0.76	0.74	0.74	**0.83**	0.81
	ARI	0.77	0.67	0.69	0.70	0.69	**0.81**	0.78
	ICAS	0.51	0.54	0.58	0.34	0.41	0.98	**0.32**
Reviews	Acc	0.52	0.51	**0.69**	0.56	0.58	0.54	0.65
	NMI	0.63	0.63	**0.79**	0.64	0.65	0.65	0.67
	ARI	0.47	0.48	**0.74**	0.49	0.52	0.51	0.55
	ICAS	0.58	0.58	0.55	0.64	0.53	0.97	**0.56**
Sports	Acc	0.33	0.34	0.28	0.35	0.37	0.38	**0.41**
	NMI	0.70	0.68	0.57	0.69	**0.70**	0.68	0.69
	ARI	0.47	0.46	0.31	0.47	**0.47**	0.45	0.47
	ICAS	0.69	0.68	0.70	0.74	0.70	0.99	**0.67**
Classic3	Acc	0.90	0.93	0.73	0.90	0.89	0.91	**0.93**
	NMI	0.76	0.80	0.60	0.76	0.74	0.80	**0.81**
	ARI	0.75	0.81	0.49	0.75	0.72	0.79	**0.81**
	ICAS	0.30	0.34	**0.27**	0.33	0.29	0.97	0.33

* A boldface number is better than the others in each row.

4.3.1. Polysemous word identification

The concept 'polysemous' is commonly used to define words that might have multiple meanings in different contexts. For instance, the use of the word 'body' with the words 'pressure' and 'operation', refers to the concept of 'medicine'. In contrast, the use of the word 'body' with the words 'wing' and 'aircraft' indicates an 'aeronautical systems'. In [9], it can be seen an

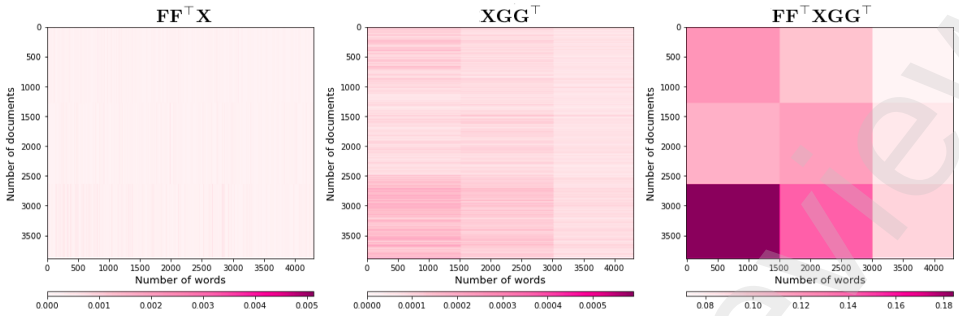

Figure 3: Result by OPNMTF with $\alpha = 1$, $\lambda = 0.03$, $\mu = 1$, $g = 3$, and $s = 3$ on **Classic3**, (Left) Clustering row data matrix ($\mathbf{FF}^\top\mathbf{X}$). (Middle) Clustering column data matrix ($\mathbf{XGG}^\top$). (Right) Co-clustering reorganization data matrix ($\mathbf{FF}^\top\mathbf{XGG}^\top$).

analysis of polysemous for **CSTR** dataset.

Figure 1 shows the structure of the summarization matrix **S**. It plays a role in identifying how Word-Cluster (WC) relates to Document-Cluster (DC). Table 5 shows an example of **S** obtained by OPNMTF on **Classic3** dataset, with $g = 3$ and $s = 3$. Document-Clusters (denoted by DC1 through DC3) are connected to Word-Clusters (denoted by WC1 through WC3). In addition, the words in WC1, such as 'layer' and 'body' are suitable for use in DC1 as a 'medical' topic and DC3 as a 'aeronautical systems' topic. Therefore, analyzing **S** allows us to easily identify polysemous words and their corresponding contexts.

DC1-WC1 in Figure 4a, DC2-WC2 in Figure 4e, and DC3-WC3 in Figure 4i contains words representing topics 'medical', 'information retrieval', 'aeronautical systems', respectively.

(a) DC1-WC1 (b) DC1-WC2 (c) DC1-WC3

(d) DC2-WC1 (e) DC2-WC2 (f) DC2-WC3

(g) DC3-WC1 (h) DC3-WC2 (i) DC3-WC3

Figure 4: Word clouds of the top 100 words for matrix **S** have been shown in Table 5 for **Classic3** dataset. The bigger word has more frequency.

Table 5: The summarization matrix $\mathbf{S}$ obtained by OPNMTF with $\alpha = 1$, $\lambda = 0.03$, and $\mu = 1$. WC1 to WC3 represent Word-Clusters 1 through 3, and DC1 to DC3 represent Document-Clusters 1 through 3.

	WC1	WC2	WC3
DC1: Medical	0.1173	0.0834	0.0856
DC2: Information retrieval	0.0935	0.0874	0.0886
DC3: Aeronautical systems	0.1956	0.1194	0.1287

4.3.2. Parameter sensitivity

In this subsection, we have two strategies for parameter sensitivities. The first plan is choosing a tuple (λ, μ) from the set $\{(10^{-9}, 10^{-9}), (10^{-9}, 10^{+9}), (10^{+9}, 10^{-9}), (10^{+9}, 10^{+9})\}$ that is defined according to the duality of very big (or small) λ with very big (or small) μ value. The second plan is choosing value α from the set $\{0, 0.5, 1\}$. The results of these strategies are shown in Figures 5 and 6 for all datasets, respectively. There is a relation between orthogonality parameters and the variance of evaluation measures for each dataset. In Figure 5, the performance of evaluation measures have a high variance when $(\lambda, \mu) \in \{(10^{-9}, 10^{-9}), (10^{-9}, 10^{+9})\}$, while the performance mentioned has low variance when $(\lambda, \mu) \in \{(10^{+9}, 10^{-9}), (10^{+9}, 10^{+9})\}$ for **RCV1**, **Classic3**, and **Reviews** datasets. The various values of $\alpha \in \{0.5, 1, 2\}$ are almost the same performance for all datasets in Figure 6.

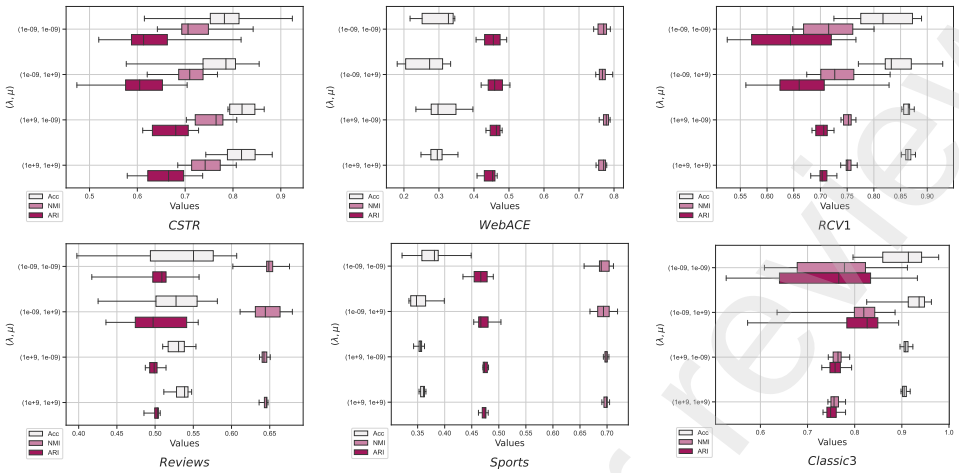

Figure 5: Box plots of measures (Acc, NMI, ARI) for six datasets are introduced in Table 3, based on 100 iterations of OPNMTF with $\alpha = 0.5$ and $(\lambda, \mu) \in \{(10^{-9}, 10^{-9}), (10^{-9}, 10^{+9}), (10^{+9}, 10^{-9}), (10^{+9}, 10^{+9})\}$.

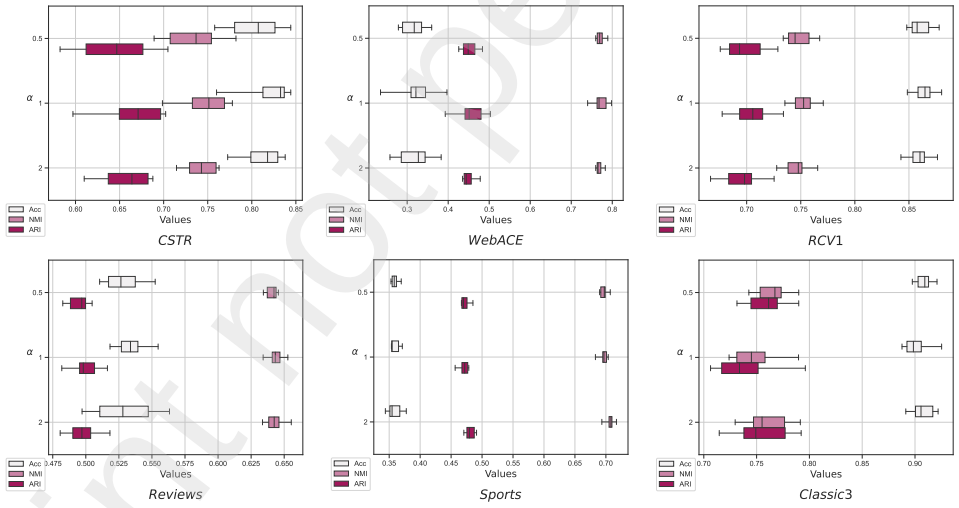

Figure 6: Box plots of measures (Acc, NMI, ARI) for six datasets are introduced in Table 3, based on 100 iterations of OPNMTF with $\lambda = 0.03$, $\mu = 1$, and $\alpha \in \{0, 0.5, 1\}$.

4.3.3. Orthogonality $\mathbf{F}$ *and* $\mathbf{G}$

The orthogonality measure for $\mathbf{F}$ and $\mathbf{G}$, computed by $\|\mathbf{F}^\top\mathbf{F} - \mathbf{I}_g\|$ and $\|\mathbf{G}^\top\mathbf{G} - \mathbf{I}_s\|$, as the iterations proceed. Figure 7 shows orthogonality mea-

sures based on extreme value of (λ, μ) for all datasets.

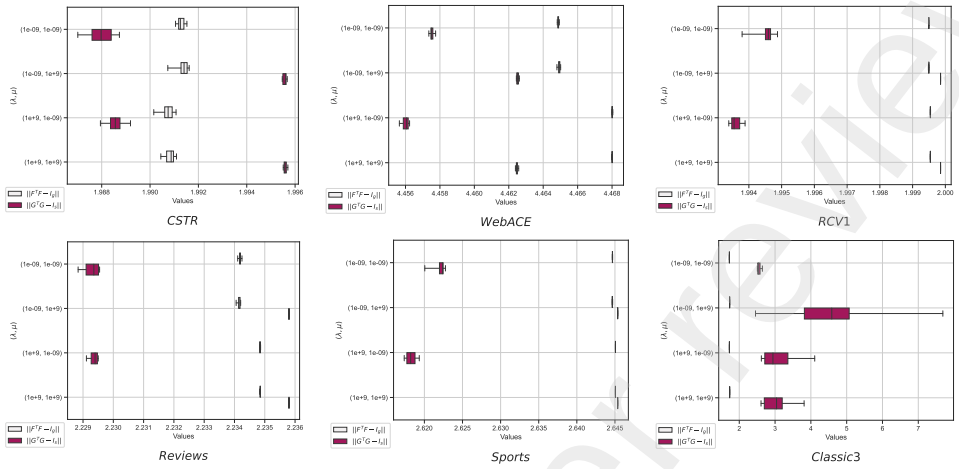

Figure 7: Box plots of orthogonality measures for $\mathbf{F}$ and $\mathbf{G}$ in six datasets are introduced in Table 3, based on 100 iterations of OPNMTF with $\alpha = 0.5$ and $(\lambda, \mu) \in \{(10^{-9}, 10^{-9}), (10^{-9}, 10^{+9}), (10^{+9}, 10^{-9}), (10^{+9}, 10^{+9})\}$.

4.3.4. Runtime algorithms

Figure 8 shows the running time list of different co-clustering algorithms in log seconds. As the number of documents and words increases, the efficacy of various co-clustering algorithms changes. For example, the proposed algorithms have a higher speed for the large datasets **RCV1** and **Sports**.

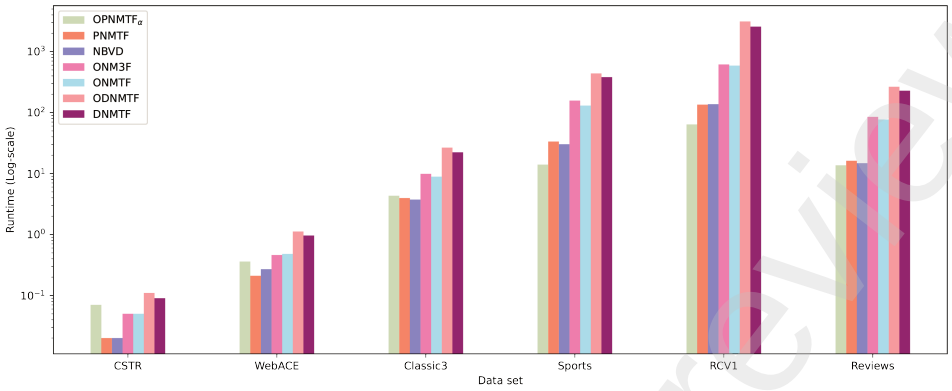

Figure 8: Runtime bar charts of different co-clustering algorithms for six datasets are reported in Table 3, based on 100 iterations for each algorithm.

5. Conclusion

The OPNMTF algorithm uses optimization techniques based on α-divergence to generate an orthogonal factor matrix as part of the objective function, rather than relying on additional constraints. This approach has been tested and is effective in several real-world datasets involving text data. In this method, the values of the parameters λ and μ can be adjusted to achieve different levels of orthogonality. However, there is no set rule for determining the best values for these parameters when using the algorithm for co-clustering. In future work, it is possible to investigate more efficient algorithms for orthogonal non-negative matrix factorization problems that incorporate penalty terms.

Appendix A. Notation

Notation	Usage
$i = 1, \ldots, n$	n is the number of rows
$j = 1, \ldots, m$	m is the number of columns
$k = 1, \ldots, g$	g is the number of row clusters
$h = 1, \ldots, s$	s is the number of column clusters
$\mathbb{R}_{+}^{n \times m}$	Real space $n \times m$ with non-negative values
$\mathbb{R}_{\{0,1\}}^{g \times g}$	Real space $g \times g$ with values zero and one
$\mathbb{R}_{1}^{g \times g}$	Real space $g \times g$ with all values one
$\mathbf{X} = (\mathrm{X}_{ij}) \in \mathbb{R}_{+}^{n \times m}$	Data matrix
$\mathbf{F} = (\mathrm{F}_{ik}) \in \mathbb{R}_{+}^{n \times g}$	Row cluster matrix
$\mathbf{S} = (\mathrm{S}_{kh}) \in \mathbb{R}_{+}^{g \times s}$	Summary co-cluster matrix
$\mathbf{G} = (\mathrm{G}_{jh}) \in \mathbb{R}_{+}^{m \times s}$	Column cluster matrix
$\mathbf{FSG}^{\top} = ([\mathbf{FSG}^{\top}]_{ij}) \in \mathbb{R}_{+}^{n \times m}$	Approximation matrix
$\mathbf{E}_{nm} \in \mathbb{R}_{1}^{n \times m}$	A matrix of ones with size $n \times m$
$\mathbf{I}_g = (I_{kk}) \in \mathbb{R}_{\{0,1\}}^{g \times g}$	Identity matrix with size $g \times g$
$\mathbf{I}_s = (I_{hh}) \in \mathbb{R}_{\{0,1\}}^{s \times s}$	Identity matrix with size $s \times s$
$\mathbf{D}_{\mathbf{F}} = \mathrm{diag}(\mathbf{1}_n^{\top}\mathbf{F})$	Diagonal matrix normalization for $\mathbf{F}$
$\mathbf{D}_{\mathbf{G}} = \mathrm{diag}(\mathbf{1}_m^{\top}\mathbf{G})$	Diagonal matrix normalization for $\mathbf{G}$
$D_{\alpha}(\mathbf{X} \Vert \mathbf{FSG}^{\top})$	α-Divergence between $\mathbf{X}$ and $\mathbf{FSG}^{\top}$
ξ_{OPNMTF}	Objective function for OPNMTF with α-divergence
λ	Orthogonality parameter row in OPNMTF
μ	Orthogonality parameter column in OPNMTF
α	Parameter α-divergence
t	Iteration index
N	Maximum number of iterations
$\mathcal{C}_k$	The prediction k-th cluster
$\mathcal{L}_{k'}$	The true k'-th class
$\mathrm{T}(\mathcal{C}_k, \mathcal{L}_{k'})$	The proportion of correct recovered objects by the clustering algorithm
n_k	The number of rows in the cluster $\mathcal{C}_k (1 \leqslant k \leqslant g)$
$\hat{n}_{k'}$	The number of rows in the class $\mathcal{L}_{k'} (1 \leqslant k' \leqslant g)$
$n_{kk'}$	The number of rows in $\mathcal{C}_k \cap \mathcal{L}_{k'}$
n	The total number of documents
$\mathrm{sim}\,(\mathrm{centroid}_k, \mathrm{centroid}_{k'})$	Cosine similarity function between k-th and k'-th cluster centroid
$\odot$	Element-wise multiplication
$\mathbf{A} \oslash \mathbf{B} = \frac{\mathbf{A}}{\mathbf{B}}$	Element-wise division matrix $\mathbf{A}$ on $\mathbf{B}$
$[\mathbf{P}]_{+} = \max(\mathbf{P}, 0)$	Applying the non-negativity condition for matrix $\mathbf{P}$
$\mathrm{Tr}(\mathbf{P})$	Trace of matrix $\mathbf{P}$

Appendix B. Proof of Lemma 2.1

Proof. Due to $A(\mathbf{F}^{(t+1)}, \mathbf{F}^{(t)})$ is an auxiliary function for $D_\alpha(\mathbf{X}\|\mathbf{F}^{(t+1)}\mathbf{S}\mathbf{G}^\top)$, we have

$$D_\alpha(\mathbf{X}\|\mathbf{F}^{(t+1)}\mathbf{S}\mathbf{G}^\top) \overset{(6)}{\leqslant} A(\mathbf{F}^{(t+1)}, \mathbf{F}^{(t)}) \overset{(7)}{\leqslant} A(\mathbf{F}^{(t)}, \mathbf{F}^{(t)}) \overset{(5)}{=} D_\alpha(\mathbf{X}\|\mathbf{F}^{(t)}\mathbf{S}\mathbf{G}^\top).$$

By repeating the updating rule in (7), we can identify the sequence of estimates that will lead to a local minimum of the cost function, such that

$$D_\alpha(\mathbf{X}\|\mathbf{F}^{(t_{min})}\mathbf{S}\mathbf{G}^\top) \leqslant \cdots \leqslant D_\alpha(\mathbf{X}\|\mathbf{F}^{(t)}\mathbf{S}\mathbf{G}^\top) \leqslant \cdots \leqslant D_\alpha(\mathbf{X}\|\mathbf{F}^{(0)}\mathbf{S}\mathbf{G}^\top).$$

□

Appendix C. Proof of Lemma 3.1

Proof. We need to show that the auxiliary function $A(\mathbf{F}, \mathbf{F}^{(t)})$ in (9) satisfies the following two conditions:

(i) $A(\mathbf{F}, \mathbf{F}) \overset{(a)}{=} \xi_{\text{OPNMTF}}$,

(ii) $A(\mathbf{F}, \mathbf{F}^{(t)}) \overset{(9)}{=} \sum_{i,j,k,h} X_{ij} Q_{ijkh}^{\mathbf{F}^{(t)}} f\left(\frac{F_{ik}S_{kh}G_{hj}^\top}{X_{ij}Q_{ijkh}^{\mathbf{F}^{(t)}}}\right) + \lambda \sum_{i,k} Q_{ik}^{\mathbf{F}^{(t)}} f\left(\frac{F_{ki}^\top F_{ik}}{Q_{ik}^{\mathbf{F}^{(t)}}}\right) +$

$\text{Const}_{\mathbf{G}} \overset{(a),(b),(c),(d)}{\geq} \sum_{i,j} X_{ij} f\left(\frac{\sum_{k,h} F_{ik}S_{kh}G_{hj}^\top}{X_{ij}}\right) + \lambda \sum_k f\left(\frac{\sum_i F_{ki}^{(t)\top} F_{ik}^{(t)}}{I_{kk}}\right) +$

$\text{Const}_{\mathbf{G}} \overset{(2),(8),(d)}{\geq} \xi_{\text{OPNMTF}}$.

These two conditions are proved by

(a) $\sum_{k,h} Q_{ijkh}^{\mathbf{F}^{(t)}} = \frac{\sum_{k,h} F_{ik}^{(t)} S_{kh} G_{hj}^\top}{\sum_{k',h'} F_{ik'}^{(t)} S_{k'h'} G_{h'j}^\top} = \frac{[\mathbf{F}^{(t)}\mathbf{S}\mathbf{G}^\top]_{ij}}{[\mathbf{F}^{(t)}\mathbf{S}\mathbf{G}^\top]_{ij}} = 1,$

(b) $\sum_i Q_{ik}^{\mathbf{F}^{(t)}} = \dfrac{\sum_i F_{ki}^{(t)\top} F_{ik}^{(t)}}{\sum_{i'} F_{ki'}^{(t)\top} F_{i'k}^{(t)}} = \dfrac{[\mathbf{F}^{(t)\top}\mathbf{F}^{(t)}]_{kk}}{[\mathbf{F}^{(t)\top}\mathbf{F}^{(t)}]_{kk}} = 1.$

(c) Because of the convexity of f in (3) and the use of Jensen's inequality,

- $\sum_{i,j,k,h} X_{ij} Q_{ijkh}^{\mathbf{F}^{(t)}} f\left(\dfrac{F_{ik}S_{kh}G_{hj}^\top}{X_{ij}Q_{ijkh}^{\mathbf{F}^{(t)}}}\right) \geq \sum_{i,j} X_{ij} f\left(\dfrac{\sum_{k,h} F_{ik}S_{kh}G_{hj}^\top}{X_{ij}}\right),$
- $\sum_{i,k} Q_{ik}^{\mathbf{F}^{(t)}} f\left(\dfrac{F_{ki}^\top F_{ik}}{Q_{ik}^{\mathbf{F}^{(t)}}}\right) \geq \sum_k f\left(\dfrac{\sum_i F_{ki}^{(t)\top} F_{ik}^{(t)}}{I_{kk}}\right).$

(d) $\text{Const}_{\mathbf{G}} \geq \mu D_\alpha(\mathbf{I}_s \| \mathbf{G}^\top \mathbf{G})$.

□

Appendix D. Proof of Theorem 3.4

Proof. Note that

$$\frac{\partial f(y)}{\partial y} = \frac{1}{\alpha}\left(1 - y^{-\alpha}\right), \tag{D.1}$$

the minimum of (9) is resolved by gradient equals zero using (D.1):

$$\frac{\partial A(\mathbf{F}, \mathbf{F}^{(t)})}{\partial F_{ik}} = \frac{1}{\alpha}\sum_{h,j} S_{kh}G_{hj}^\top \left(1 - (\frac{F_{ik}S_{kh}G_{hj}^\top}{X_{ij}Q_{ijkh}^{\mathbf{F}^{(t)}}})^{-\alpha}\right) + \frac{1}{\alpha} 2\lambda \sum_k F_{ik}\left(1 - (\frac{F_{ki}^\top F_{ik}}{Q_{ik}^{\mathbf{F}^{(t)}}})^{-\alpha}\right) = 0,$$

which gives rise to

$$\left(\frac{F_{ik}}{F_{ik}^{(t)}}\right)^\alpha = \frac{\sum_{h,j} S_{kh}G_{hj}^\top \left(\dfrac{X_{ij}}{\sum_{k',h'} F_{ik'}^{(t)} S_{k'h'} G_{h'j}^\top}\right)^\alpha + 2\lambda \sum_k F_{ik}\left(\dfrac{I_{kk}}{[\mathbf{F}^\top\mathbf{F}]_{kk}}\right)^\alpha}{\sum_{h,j} S_{kh}G_{hj}^\top + 2\lambda \sum_k F_{ik}}.$$

Therefore, we suggest the element-wise updating rule for F_{ik} is:

$$F_{ik}^{(t+1)} \leftarrow F_{ik}^{(t)} \left(\frac{\sum_{h,j} S_{kh} G_{hj}^{\top} \left(\frac{X_{ij}}{[\mathbf{FSG}^{\top}]_{ij}} \right)^{\alpha} + 2\lambda \sum_{k} F_{ik} \left(\frac{I_{kk}}{[\mathbf{F}^{\top}\mathbf{F}]_{kk}} \right)^{\alpha}}{\sum_{j} [\mathbf{SG}^{\top}]_{kj} + 2\lambda \sum_{k} F_{ik}} \right)^{\frac{1}{\alpha}},$$

and suggest the matrix-wise updating rule for $\mathbf{F}$

$$\mathbf{F}^{(t+1)} \leftarrow \mathbf{F}^{(t)} \odot \left[\left(\frac{(\mathbf{X} \oslash \mathbf{FSG}^{\top})^{\alpha} \mathbf{GS}^{\top} + 2\lambda \mathbf{F}(\mathbf{I}_g \oslash \mathbf{F}^{\top}\mathbf{F})^{\alpha}}{\mathbf{E}_{nm} \mathbf{GS}^{\top} + 2\lambda \mathbf{FI}_g} \right)^{\frac{1}{\alpha}} \right]_{+},$$

that is identical to (12).

similarity, the updating rule (13) and (14) are determined by solving $\frac{\partial A(\mathbf{G}, \mathbf{G}^{(t)})}{\partial G_{jh}}$ and $\frac{\partial A(\mathbf{S}, \mathbf{S}^{(t)})}{\partial S_{kh}}$, where $A(\mathbf{G}, \mathbf{G}^{(t)})$ and $A(\mathbf{S}, \mathbf{S}^{(t)})$ are given in (10) and (11). □

References

[1] H. Shan, A. Banerjee, Generalized probabilistic matrix factorizations for collaborative filtering, Proceedings - IEEE International Conference on Data Mining, ICDM (2010) 1025 – 1030.

[2] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining, Journal for Language Technology and Computational Linguistics 20 (2005) 19–62.

[3] T. Hofmann, J. Puzicha, M. I. Jordan, Learning from dyadic data, Advances in Neural Information Processing Systems (1999) 466–472.

[4] B. Long, Z. Zhang, P. S. Yu, Co-clustering by block value decom-

position, Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (2005) 635–640.

[5] N. Del Buono, G. Pio, Non-negative matrix tri-factorization for co-clustering: an analysis of the block matrix, Information Sciences 301 (2015) 13–26.

[6] J. A. Hartigan, Direct clustering of a data matrix, Journal of the American Statistical Association 67 (1972) 123–129.

[7] G. Govaert, M. Nadif, Co-clustering: models, algorithms and applications, John Wiley & Sons, 2013.

[8] C. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix tri-factorizations for clustering, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006) 126–135.

[9] J. Yoo, S. Choi, Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds, Information Processing & Management 46 (2010) 559–570.

[10] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.

[11] D. Seung, L. Lee, Algorithms for non-negative matrix factorization, Advances in neural information processing systems 13 (2001) 556–562.

[12] V. P. Pauca, J. Piper, R. J. Plemmons, Nonnegative matrix factorization for spectral data analysis, Linear Algebra and its Applications 416 (2006) 29–47.

[13] Y.-X. Wang, Y.-J. Zhang, Nonnegative matrix factorization: A comprehensive review, IEEE Transactions on Knowledge and Data Engineering 25 (2012) 1336–1353.

[14] T. Li, C. C. Ding, Nonnegative matrix factorizations for clustering: A survey, Data Clustering (2018) 149–176.

[15] L. Labiod, M. Nadif, Co-clustering under nonnegative matrix tri-factorization, International Conference on Neural Information Processing 7063 LNCS (2011) 709–717.

[16] S. Wang, A. Huang, Penalized nonnegative matrix tri-factorization for co-clustering, Expert Systems with Applications 78 (2017) 64–73.

[17] A. Cichocki, H. Lee, Y.-D. Kim, S. Choi, Non-negative matrix factorization with α-divergence, Pattern Recognition Letters 29 (2008) 1433–1440.

[18] A. Cichocki, S.-i. Amari, R. Zdunek, R. Kompass, G. Hori, Z. He, Extended smart algorithms for non-negative matrix factorization, International Conference on Artificial Intelligence and Soft Computing (2006) 548–562.

[19] A. Cichocki, R. Zdunek, S.-i. Amari, Csiszar's divergences for non-negative matrix factorization: Family of new algorithms, International Conference on Independent Component Analysis and Signal Separation (2006) 32–39.

[20] S. Sra, I. Dhillon, Generalized nonnegative matrix approximations with Bregman divergences, Advances in Neural Information Processing Systems 18 (2005).

[21] A. Cichocki, R. Zdunek, S.-i. Amari, New algorithms for non-negative matrix factorization in applications to blind source separation, IEEE International Conference on Acoustics Speech and Signal Processing Proceedings 5 (2006) V–V.

[22] S.-i. Amari, Differential-geometrical methods in statistics, Springer Science & Business Media, 2012.

[23] H. W. Kuhn, Nonlinear programming: a historical view, Traces and Emergence of Nonlinear Programming (2014) 393–414.

[24] C. Laclau, M. Nadif, Hard and fuzzy diagonal co-clustering for document-term partitioning, Neurocomputing 193 (2016) 133–147.

[25] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research 3 (2002) 583–617.

[26] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.

[27] T. Li, A general model for clustering binary data, Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (2005) 188–197.

[28] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore, Webace: A web agent for document categorization and exploration, Proceedings of the Second International Conference on Autonomous Agents (1998) 408–415.

[29] D. Cai, X. He, Manifold adaptive experimental design for text categorization, IEEE Transactions on Knowledge and Data Engineering 24 (2011) 707–719.

[30] G. Karypis, Cluto-a clustering toolkit, University of Minnesota, Technical Report (2002).

[31] I. S. Dhillon, S. Mallela, D. S. Modha, Information-theoretic coclustering, in: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 89–98.