Joint Local and Global Consistency on Interdocument and Interword Relationships for Co-Clustering

Bing-Kun Bao, Weiqing Min, Teng Li, and Changsheng Xu, *Fellow, IEEE*

***Abstract*—Co-clustering has recently received a lot of attention due to its effectiveness in simultaneously partitioning words and documents by exploiting the relationships between them. However, most of the existing co-clustering methods neglect or only partially reveal the interword and interdocument relationships. To fully utilize those relationships, the local and global consistencies on both word and document spaces need to be considered, respectively. Local consistency indicates that the label of a word/document can be predicted from its neighbors, while global consistency enforces a smoothness constraint on words/documents labels over the whole data manifold. In this paper, we propose a novel co-clustering method, called co-clustering via local and global consistency, to not only make use of the relationship between word and document, but also jointly explore the local and global consistency on both word and document spaces, respectively. The proposed method has the following characteristics: 1) the word-document relationships is modeled by following information-theoretic co-clustering (ITCC); 2) the local consistency on both interword and interdocument relationships is revealed by a local predictor; and 3) the global consistency on both interword and interdocument relationships is explored by a global smoothness regularization. All the fitting errors from these three-folds are finally integrated together to formulate an objective function, which is iteratively optimized by a convergence provable updating procedure. The extensive experiments on two benchmark document datasets validate the effectiveness of the proposed co-clustering method.**

***Index Terms*—Co-clustering, information theory, local and global learning.**

I. Introduction

CO-CLUSTERING, a recent technique to simultaneously cluster both words and documents, has received significant attention [1]–[4]. Unlike one-side clustering which just groups similar words (or similar documents), co-clustering makes use of the relationships between words and documents, referred to word-document relationships in this paper, and simultaneously groups them into word clusters and document clusters.

According to the way to explore word-document relationships, co-clustering can be divided into three categories, i.e., bipartite graph partition-based, information theory-based and matrix factorization-based. The first category models word-document relationships by the edges between word vertex and document vertex, and the objective of co-clustering is to find minimum cut vertex partitions in the bipartite graph between words and documents [1], [4]–[6]. The second one considers word-document contingency table as an empirical joint probability distribution of two random variables, respectively, from word and document spaces, and the co-clustering is to pursue the mutual information between clustered random variables to approximate to that between the original random variables [2], [7]–[10]. And for the third one, matrix factorization-based co-clustering methods co-clusters both words and documents by data reconstruction through nonnegative matrix factorization on word-document matrix [3], [11], [12]. Lots of variants of these three categories of co-clustering methods exist, but most of them focus on building different strategies to model word-document relationship, yet neglect the relationships among words and those among documents, called interword and interdocument relationships in this paper.

It is obvious that the interword and interdocument relationships cannot be ignored in the co-clustering process, since words/documents of the same cluster are expected to be more similar than others from different clusters. Fig. 1 shows an example of the necessity of incorporating interword and interdocument relationships. If considering word-document relationship only, the resulted word clusters are $\{x_1, x_2\}$, $\{x_4, x_5\}$, $\{x_3, x_6, x_7, x_8, x_9\}$ and document clusters are $\{y_1\}$, $\{y_4, y_5\}$, $\{y_2, y_3, y_6, y_7, y_8\}$. However, when incorporating interword and interdocument relationships, x_3 should be clustered together with $\{x_4, x_5\}$, y_2 should be together with $\{y_1\}$, and y_3 should be grouped into cluster $\{y_4, y_5\}$.

To model the interword and interdocument relationships, two factors can be considered: local consistency and global consistency. Local consistency indicates that the assignment function of a document/word need be consistent with the predictor of its neighbors, while global consistency enforces a

Manuscript received April 27, 2013; revised November 22, 2013, March 13, 2014, and April 10, 2014; accepted April 10, 2014. Date of publication June 26, 2014; date of current version December 15, 2014. This work was supported in part by National Program on Key Basic Research Project 973 Program, Project 2012CB316304, in part by the National Natural Science Foundation of China under Grant 61201374, Grant 61300056, and Grant 61225009, in part by China Post-Doctoral Science Foundation under Grant 2013T60196, and in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 201306282. This paper was recommended by Associate Editor J. Basak.

B.-K. Bao, W. Min, and C. Xu are with the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing 100191, China (e-mail: bkbao@nlpr.ia.ac.cn; wqmin@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn).

T. Li is with College of Electrical Engineering and Automation, Anhui University, Hefei 230601, China (e-mail: tenglwy@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2317514

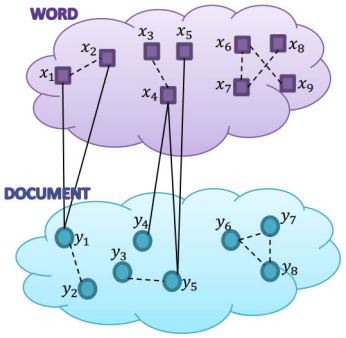

Fig. 1. Illustration of the necessity of incorporating interword and interdocument relationships. The dark edges between word and document vertex indicate word-document relationship, and black dotted edges between word/document vertex indicate interword/interdocument relationships.

smoothness constraint on documents/words assignments over the whole dataset. Fig. 2, showing two examples on toy data in single side clustering, illustrates the necessity of both local and global consistency on either word or document spaces. The issue of pure global consistency ignores the effect from neighbors which are more related to the particular test data, shown in Fig. 2(b), while pure local consistency neglects the information of the whole data manifold, shown in Fig. 2(f). Several work dedicated to this objective. Wu *et al.* [13] incorporated the word similarities and document similarities into co-clustering. Zhang *et al.* [14] investigated the local regions in word and document spaces and applied local linear regression to capture the this kind of relationships. However, these methods just enforce a smoothness constraint on cluster assignments based on only the whole data manifold [13] or only the local structure [14], and do not fully reveal the interword and interdocument relationships.

In this paper, we propose a novel co-clustering algorithm named co-clustering via local and global consistency (CCLGC) to fully explore the word-document relationship as well as the interword and interdocument relationships in terms of both local regions and global data manifold. The basic idea is to incorporate the combination of local and global regularization on both document and word spaces into information theoretic co-clustering frame. Specifically, the word-document relationships are revealed by following ITCC, which views word-document contingency table as an empirical joint probability distribution of two random variables, and minimizes the loss in mutual information between the original random variables and that between clustered random variables. The local consistency on both interdocument and interword relationships are discovered by a local predictor, which is constructed as linear regression and kernel regression, respectively. And their global consistency is revealed by a smoothness constraint on the cluster assignments over the whole data manifold. All the fitting errors from these threefolds are finally integrated together to formulate an objective function, which is iteratively optimized by a convergence provable updating procedure. The complexity analysis shows the effectiveness of our proposed CCLGC, and experimental evaluations on two benchmark document datasets show that CCLGC performs better than many state-of-art clustering and co-clustering methods.

The rest of this paper is organized as follows: in Section II, a briefly review of the related work is given. Section III gives the notation used in this paper. In Section IV, we introduce our model in detail. The algorithm and its complexity analysis are described in Section V. The experimental results on several datasets are reported in Section VI, followed by conclusion in Section VII.

II. Related Work

As an effective method to simultaneously cluster two correlated yet heterogeneous data, co-clustering has recently received plenty of attention in several research areas, such as words and documents in text mining [2], [16], genes and experimental conditions in bioinformatics [4], [17], tokens and contexts in natural language processing [18]–[20]. In the following, we will firstly briefly review the three categories of co-clustering methods, and then introduce the related work on utilizing the local and global information in label prediction and cluster assignment.

In bipartite graph partition based co-clustering methods, Dhillon *et al.* [1] proposed bipartite spectral graph partitioning (BSGP) method by finding minimum cut vertex partitions in a bipartite graph between words and documents. To extend the two data types into multiple ones, Gao *et al.* [5] proposed consistent bipartite graph co-partitioning by using high-order co-clustering on star structure, in which there is a central domain that connects any other domains to form a star structure of the interrelationship. However, the limitation of graph partition-based methods is that each word cluster needs to be associated with at least one document cluster, and the number of word clusters needs to be equal to that of document clusters. The information theory based co-clustering methods are proposed to overcome this limitation. Dhillon *et al.* [2] proposed ITCC approach by finding a pair of maps from words to word clusters and from documents to document clusters with minimum mutual information loss on words and documents. Later, Banerjee *et al.* [21] suggested a generalization maximum entropy co-clustering approach by appealing to Bregman information principle. In this kind of methods, words and documents are regarded as the instances of two random variables, and the co-clustering can be viewed as the process of compressing the associated random variables. The third type of co-clustering methods is matrix factorization-based. Block value decomposition (BVD) was proposed to factorize the original data matrix into three components, the row coefficient matrix, the block value matrix, and the column-coefficient matrix [11]. Similarly, Ding *et al.* [12] proposed nonnegative matrix tri-factorization to simultaneously cluster rows and columns of the input data matrix.

Many research studies utilize the local and global information in label prediction and cluster assignment. Zhou *et al.* [22] proposed a novel semi-supervised learning method with local and global consistency. Wu *et al.* [15] used local learning in an unsupervised manner, and formulated a clustering objective function with local linear regularization. Sun *et al.* [23]

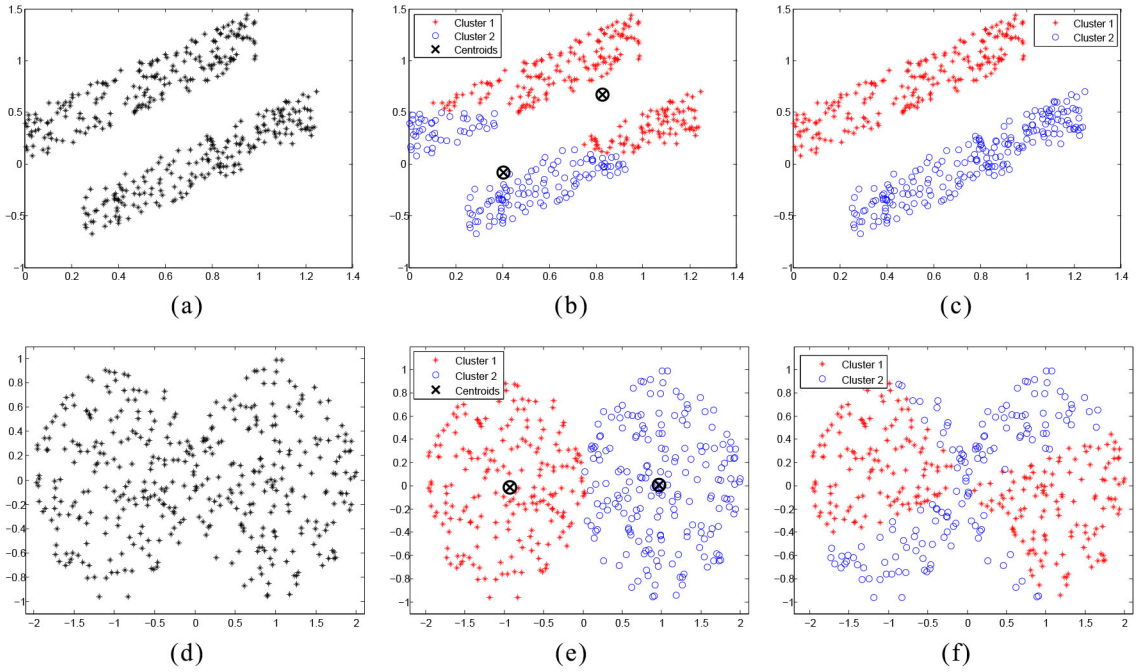

Fig. 2. (a) and (d) Two toy data distributions. (b) and (e) Clustering results with global consistency (K-means) on (a) and (d), respectively. (c) and (f) Clustering results with local consistency (LLCA [15]) on (a) and (d), respectively.

measured the fitting loss with sum of absolute error instead of sum of squared error for noise and outliers. Wang *et al.* [24] introduced both local and global regularization into clustering and achieved good performance. Several co-clustering methods are aware of the interword and interdocument relationships, but not fully discover them. Wu *et al.* [13] incorporated the word similarities and document similarities into co-clustering. Zhang *et al.* [14] investigated the local regions in word and document spaces and applied local linear regression to capture the interword and interdocument relationships.

In our work, for modeling word-document relationship, we choose information theory based method as it has no constraint on pairwise association between all the word and document clusters, and also is very effective. For modeling interword and interdocument relationships, we investigate both local and global consistency within word and document spaces, and incorporate them into information theory based co-clustering frame.

III. Notation

In this section, we give a brief introduction about the notation. Sets such as $\mathcal{A}$ are denoted by upper case letters in Euler script. Matrices are denoted using upper case bold letters, for example, $\mathbf{Y}$. Vectors are denoted by lower case bold letters, for example, $\mathbf{c} = [c_1, c_2, \ldots, c_n]^T$. Upper case letters like Y are used to denote random variables. Subscript i in c_i indicates the i-th element in vector $\mathbf{c}$, while superscript in N^i indicates the i-th neighborhood set. Subscript X and Y refer to word and document spaces, respectively, and superscript L and G refer to local consistency and global consistency. Please refer to Table I for the key notations.

TABLE I
List of Key Notations

Notation	Description
$\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$	Word set
$\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$	Document set
$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] =$ $\mathbf{Y} = [\mathbf{y}_1^T; \cdots; \mathbf{y}_m^T]$ $\in \Re^{m \times n}$	Nonnegative data matrix with the i-th column $\mathbf{x}_i$ as word sample and the j-th row $\mathbf{y}_j^T$ as document sample
$\hat{\mathcal{X}} = \{\hat{\mathcal{X}}_1, \cdots, \hat{\mathcal{X}}_{\hat{n}}\}$	Word cluster set
$\hat{\mathcal{Y}} = \{\hat{\mathcal{Y}}_1, \cdots, \hat{\mathcal{Y}}_{\hat{m}}\}$	Document cluster set
$X, Y, \hat{X}, \hat{Y}$	Random variables taking values in $\mathcal{X}, \mathcal{Y}, \hat{\mathcal{X}}$ and $\hat{\mathcal{X}}$ respectively
$\mathbf{c} = [c_1, \cdots, c_n]^T$	c_i is the assigned cluster's index for $\mathbf{x}_i$
$\mathbf{d} = [d_1, \cdots, d_m]^T$	d_i is the assigned cluster's index for $\mathbf{y}_i$
$\mathcal{N}^i$	Set of $\mathbf{x}_i$ together with its k-nearest neighbors
$\mathcal{M}^i$	Set of $\mathbf{y}_i$ together with its k-nearest neighbors
n^i, m^i	Cardinalities of $\mathcal{N}^i$ and $\mathcal{M}^i$
$\mathcal{A}^i = \{k \vert \mathbf{x}_k \in \mathcal{N}^i\}$	Set containing indices of samples in $\mathcal{N}^i$
$\mathcal{B}^i = \{k \vert \mathbf{y}_k \in \mathcal{M}^i\}$	Set containing indices of samples in $\mathcal{M}^i$
$\mathbf{N}^i = [(x_{\mathcal{A}^i})^T]$ $\in \Re^{n^i \times m}$	Matrix of word samples from $\mathcal{N}^i$
$\mathbf{M}^i = [\mathbf{y}_{\mathcal{B}^i}]$ $\in \Re^{n \times m^i}$	Matrix of document samples from $\mathcal{M}^i$

IV. Problem Formulation

The goal of co-clustering is to simultaneously group word set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ into $\hat{n}$ disjoint word clusters $\{\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2, \ldots, \hat{\mathcal{X}}_{\hat{n}}\}$ and document set $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m\}$ into $\hat{m}$ disjoint document clusters set $\{\hat{\mathcal{Y}}_1, \hat{\mathcal{Y}}_2, \ldots, \hat{\mathcal{Y}}_{\hat{m}}\}$, respectively. As we mentioned before, the main idea behind our algorithm

is to simultaneously cluster words and documents by fully making use of word-document relationship, as well as interword and interdocument relationships. Thus, there are three kinds of relationships, which exist between word and document, among word samples, and among document samples. This section successively introduces the methods of modeling these three kinds of relationships, and then formulates the overall objective function for CCLGC.

A. Modeling the Word-Document Relationship

In our work, we model relationship between word and document by following ITCC [2], which minimizes the loss in mutual information between the original random variables and that between the clustered ones. Compared with another solution of co-clustering on interrelationship, i.e., BSGP [1], ITCC has no restriction that each document cluster needs to be associated with a word cluster.

According to ITCC [2], the objective that we aim to minimize is the loss in mutual information

$$J_{XY} = I(X; Y) - I(\hat{X}; \hat{Y}) \tag{1}$$

where $I(X; Y)$ is the mutual information between X and Y, and $I(\hat{X}; \hat{Y})$ is that between $\hat{X}$ and $\hat{Y}$, which are defined as

$$I(X; Y) = \sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{x}) p(\mathbf{y}|\mathbf{x}) \log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} \tag{2}$$

$$I(\hat{X}; \hat{Y}) = \sum_{\hat{x}} \sum_{\hat{y}} p(\hat{\mathcal{X}}) p(\hat{\mathcal{Y}}|\hat{\mathcal{X}}) \log \frac{p(\hat{\mathcal{Y}}|\hat{\mathcal{X}})}{p(\hat{\mathcal{Y}})}. \tag{3}$$

After deduction, (1) can be expressed as 1) a weighted sum of the relative entropies between word distributions $p(Y|\mathbf{x})$ and word cluster prototype distributions $p(Y|\hat{x})$, referred to (4) or as 2) a weighted sum of the relative entropies between document distributions $p(X|\mathbf{y})$ and document cluster prototype distributions $p(X|\hat{\mathcal{Y}})$, referred to (5)

$$J_{XY} = \sum_{\hat{\mathcal{X}}} \sum_{\mathbf{x} \in \hat{\mathcal{X}}} p(\mathbf{x}) D(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}})) \tag{4}$$

$$J_{XY} = \sum_{\hat{\mathcal{Y}}} \sum_{\mathbf{y} \in \hat{\mathcal{Y}}} p(\mathbf{y}) D(p(X|\mathbf{y}) || q(X|\hat{\mathcal{Y}})) \tag{5}$$

where $q(\mathbf{y}|\hat{\mathcal{X}}) = p(\mathbf{y}|\hat{\mathcal{Y}}) p(\hat{\mathcal{Y}}|\hat{\mathcal{X}})$, $q(\mathbf{x}|\hat{\mathcal{Y}}) = p(\mathbf{x}|\hat{\mathcal{X}}) p(\hat{\mathcal{X}}|\hat{\mathcal{Y}})$, and $D(\cdot||\cdot)$ denotes relative entropy, also known as Kullback–Leibler (KL) divergence. Although ITCC can reveal the relationship between word and document, but it neglects the interword and interdocument relationships.

B. Modeling the Interword Relationship

It is widely accepted in single-way clustering that both local and global information could be useful in the cluster assignment [22], [24], referred to Fig. 2. This motivates us to incorporate local and global consistency on interword relationship into co-clustering: 1) local consistency: the nearby words are likely to follow the same predictor in assignment and 2) global consistency: the cluster assignment predictions should be smooth over the whole data manifold. Therefore, for interword relationship, we let every word sample iteratively spread its cluster assignment from its neighbors on one hand, and apply a smoother to smooth the predictor with respect to the intrinsic data manifold.

1) Local Consistency in Word Space: Define f_X^i be a local label predictor to estimate the clusters assigned for all the word samples in $\mathcal{N}^i$. Considering to tackle both linear and nonlinear problems, we respectively choose linear regression and kernel regression as the local predictor.

a) Linear Regression Model: We consider to fit predictor f_X^i with the following learning form:

$$f_X^i(\mathbf{x}_j) = \mathbf{x}_j^T \mathbf{w}_X^i + b_X^i \tag{6}$$

where $\mathbf{x}_j$ is a k-nearest neighbor of $\mathbf{x}_i$, that is $\mathbf{x}_j \in \mathcal{N}^i$. $\mathbf{w}_X^i \in \Re^{m \times 1}$ and $b_X^i \in \Re$ are the coefficient and intercept in f_X^i. Therefore, for the i-th subset $\mathcal{N}^i$, the target of local consistency is to pursue the estimated labels of $\mathbf{x}_i$'s neighbors to approximate their labels, that is, to minimize the fitting error on $\mathcal{N}^i$

$$J_X^L(i) = \frac{1}{n^i} \sum_{\mathbf{x}_j \in \mathcal{N}^i} \|\mathbf{x}_j^T \mathbf{w}_X^i + b_X^i - c_j\|_2^2 + \lambda_X^i \|\mathbf{w}_X^i\|_2^2 \tag{7}$$

where $\|\cdot\|_2$ is the ℓ_2-norm of a vector. The penalty term $\lambda_X^i \|\mathbf{w}_X^i\|_2^2$ is introduced to avoid overfitting [25]. Let $\mathbf{1}$ be a n^i-dimensional column vector of all ones, $J_X^L(i)$ can be rewritten as

$$J_X^L(i) = \frac{1}{n^i} \|\mathbf{N}^i \mathbf{w}_X^i + b_X^i \mathbf{1} - \mathbf{c}^i\|_2^2 + \lambda_X^i \|\mathbf{w}_X^i\|_2^2 \tag{8}$$

where $\mathbf{c}^i = \mathbf{S}^i \mathbf{c} \in \Re^{n^i}$, and $\mathbf{S}^i = [\mathbf{e}_k]^T, k \in \mathcal{A}^i$ is a selection matrix for each subset $\mathcal{N}^i$, $\mathbf{e}_k$ is a n-dimensional vector whose k-th element is one and all others are zeros.

Taking the first order partial derivatives of (7) with respective to $\mathbf{w}_X^i$ and b_X^i, we have

$$\begin{cases} \frac{\partial J_X^L(i)}{\partial \mathbf{w}_X^i} = \frac{2(\mathbf{N}^i)^T}{n^i} [\mathbf{N}^i \mathbf{w}_X^i + b_X^i \mathbf{1} - \mathbf{c}^i] + 2\lambda_X^i \mathbf{w}_X^i \\ \frac{\partial J_X^L(i)}{\partial b_X^i} = \frac{2}{n^i} [\mathbf{1}^T \mathbf{N}^i \mathbf{w}_X^i + n^i b_X^i - \mathbf{1}^T \mathbf{c}^i]. \end{cases} \tag{9}$$

Let $\frac{\partial J_X^L(i)}{\partial \mathbf{w}_X^i} = 0$ and $\frac{\partial J_X^L(i)}{\partial b_X^i} = 0$, then

$$\begin{cases} (\mathbf{w}_X^i)^\star = [(\mathbf{N}^i)^T \mathbf{E}_X \mathbf{N}^i + n^i \lambda_X^i \mathbf{I}]^{-1} (\mathbf{N}^i)^T \mathbf{E}_X \mathbf{c}^i \\ (b_X^i)^\star = \frac{\mathbf{1}}{n^i} [(\mathbf{1})^T \mathbf{c}^i - (\mathbf{1})^T \mathbf{N}^i (\mathbf{w}_X^i)^\star] \end{cases} \tag{10}$$

where $\mathbf{E}_X = \mathbf{I} - \frac{1}{n^i} \mathbf{1}\mathbf{1}^T$, and $\mathbf{I}$ is identity matrix with proper dimension.

Substituting $(\mathbf{w}_X^i)^\star$ and $(b_X^i)^\star$ into (8), we can get

$$\begin{aligned} J_X^L(i) &= \frac{1}{n^i} (\mathbf{c}^i)^T \Big[\mathbf{E}_X - \mathbf{E}_X \mathbf{N}^i \big((\mathbf{N}^i)^T \mathbf{E}_X \mathbf{N}^i \\ &\qquad + n^i \lambda_X^i \mathbf{I} \big)^{-1} (\mathbf{N}^i)^T \mathbf{E}_X \Big] \mathbf{c}^i \\ &= (\mathbf{c}^i)^T \mathbf{L}_X^i \mathbf{c}^i \\ &= \mathbf{c}^T (\mathbf{S}^i)^T \mathbf{L}_X^i \mathbf{S}^i \mathbf{c} \end{aligned}$$

where

$$\mathbf{L}_X^i = \frac{1}{n^i} \Big[\mathbf{E}_X - \mathbf{E}_X \mathbf{N}^i ((\mathbf{N}^i)^T \mathbf{E}_X \mathbf{N}^i + n^i \lambda_X^i \mathbf{I})^{-1} (\mathbf{N}^i)^T \mathbf{E}_X \Big]. \tag{11}$$

Note that the computational cost of (11) is expensive when the number of words is large, since it needs to inverse a $n \times n$ matrix. Using the Woodbury formula [26], [27], we can rewrite (11) as (12), which only involves the inverse of a $n^i \times n^i$ matrix

$$\begin{aligned}\mathbf{L}_X^i &= \frac{1}{n^i}\left[\mathbf{E}_X - \mathbf{E}_X\mathbf{N}^i((\mathbf{N}^i)^T\mathbf{E}_X\mathbf{N}^i + n^i\lambda_X^i\mathbf{I})^{-1}(\mathbf{N}^i)^T\mathbf{E}_X\right] \\ &= \frac{1}{n^i}\mathbf{E}_X\left[\mathbf{I} - \mathbf{E}_X\mathbf{N}^i\left((\mathbf{N}^i)^T\mathbf{E}_X\mathbf{N}^i + n^i\lambda_X^i\mathbf{I}\right)^{-1}(\mathbf{N}^i)^T\mathbf{E}_X\right]\mathbf{E}_X \\ &= \frac{1}{n^i}\mathbf{E}_X\Big[\mathbf{I} - \mathbf{I}\mathbf{E}_X\mathbf{N}^i\left((\mathbf{N}^i)^T\mathbf{E}_X\mathbf{I}\mathbf{E}_X\mathbf{N}^i + n^i\lambda_X^i\mathbf{I}\right)^{-1}(\mathbf{N}^i)^T \\ &\qquad \mathbf{E}_X\mathbf{I}\Big]\mathbf{E}_X \\ &= \frac{1}{n^i}\mathbf{E}_X\left(\mathbf{I} + \frac{1}{n^i\lambda_X^i}\mathbf{E}_X\mathbf{N}^i(\mathbf{N}^i)^T\mathbf{E}_X\right)^{-1}\mathbf{E}_X \\ &= \lambda_X^i\mathbf{E}_X\left(n^i\lambda_X^i\mathbf{I} + \mathbf{E}_X\mathbf{N}^i(\mathbf{N}^i)^T\mathbf{E}_X\right)^{-1}\mathbf{E}_X. \end{aligned} \tag{12}$$

Therefore, the overall objective function on interword relationship with local consistency should minimize the combination of all the local label predictors

$$\begin{aligned} J_X^L &= \sum_{i=1}^n J_X^L(i) = \sum_{i=1}^n (\mathbf{c}^i)^T\mathbf{L}_X^i\mathbf{c}^i \\ &= \sum_{i=1}^n [\mathbf{c}^T(\mathbf{S}^i)^T\mathbf{L}_X^i\mathbf{S}^i\mathbf{c}] \\ &= \mathbf{c}^T\sum_{i=1}^n [(\mathbf{S}^i)^T\mathbf{L}_X^i\mathbf{S}^i]\mathbf{c}. \end{aligned} \tag{13}$$

Let $\mathbf{L}_X = \sum_{i=1}^n [(\mathbf{S}^i)^T\mathbf{L}_X^i\mathbf{S}^i]$, then

$$J_X^L = \mathbf{c}^T\mathbf{L}_X\mathbf{c} \tag{14}$$

b) Kernel Regression Model: The linear projector can only handle the linear problems. In order to solve nonlinear problem, we extend to choose kernel regression model as our local label predictor. Let ϕ:$\Re^m \to \mathcal{F}_X$ be a mapping to map the word $\mathbf{x}$ into a high dimensional space $\mathcal{F}_X$, in which we can get a linear label predictor

$$f_X^i(\mathbf{x}_j) = \phi(\mathbf{x}_j)\tilde{\mathbf{w}}_X^i + b_X^i \tag{15}$$

where $\mathbf{x}_j$ is a k-nearest neighbor of $\mathbf{x}_i$, that is $\mathbf{x}_j \in \mathcal{N}^i$, $\tilde{\mathbf{w}}_X^i$ and b^i are the parameters of f_X^i. Since $\tilde{\mathbf{w}}_X^i$ lies in the span of $\{\phi(\mathbf{x}_j)|\mathbf{x}_j \in \mathcal{N}^i\}$, there exists a coefficient vector $\mathbf{v}^i \in \Re^{n^i}$ such that

$$\tilde{w}_X^i = (\phi(\mathbf{N}^i))^T\mathbf{v}_X^i. \tag{16}$$

Then we have

$$f_X^i(\mathbf{N}^i) = \phi(\mathbf{N}^i)(\phi(\mathbf{N}^i))^T\mathbf{v}_X^i + b_X^i\mathbf{1}. \tag{17}$$

Then, we need to minimize the loss error on the i-th subset $\mathcal{N}^i$

$$\begin{aligned} J_X^L(i) = \frac{1}{n^i}\|\phi(\mathbf{N}^i)(\phi(\mathbf{N}^i))^T\mathbf{v}_X^i + b_X^i\mathbf{1} - \mathbf{c}^i\|^2 \\ +\lambda_X^i\|(\phi(\mathbf{N}^i))^T\mathbf{v}_X^i\|^2 \end{aligned} \tag{18}$$

where $\mathbf{1}$ is vector of all ones with the dimension equal to n^i. Similarly to (7), $\lambda_X^i\|(\phi(\mathbf{N}^i))^T\mathbf{v}_X^i\|_F^2$ is a penalty term.

Define $\mathbf{K}_X^i \in \Re^{n^i \times n^i}$ be the kernel matrix for $\mathbf{N}^i$ with its (i,j)-th entry $\mathbf{K}_X^i(k,j) = \mathbf{K}_X(\mathbf{x}_{i_k}, \mathbf{x}_{i_j})$, where $\mathbf{x}_{i_j}$ is the j-th element of $\mathbf{x}^i$. Taking the first order partial derivatives of (18) with respective to $\mathbf{v}_X^i$ and b_X^i, we have

$$\begin{cases} \frac{\partial J_X^L(i)}{\partial \mathbf{v}_X^i} = \frac{2(\mathbf{K}_X^i)^T}{n^i}\left[\mathbf{K}_X^i\mathbf{v}_X^i + b_X^i\mathbf{1} - \mathbf{c}^i\right] + 2\lambda_X^i\mathbf{K}_X^i\mathbf{v}_X^i \\ \frac{\partial J_X^L(i)}{\partial b_X^i} = \frac{2}{n^i}\left[\mathbf{1}^T\mathbf{K}_X^i\mathbf{v}_X^i + n^ib_X^i - \mathbf{1}^T\mathbf{c}^i\right]. \end{cases} \tag{19}$$

Let $\frac{\partial J_X^L(i)}{\partial \mathbf{v}_X^i} = 0$ and $\frac{\partial J_X^L(i)}{\partial b_X^i} = 0$ then

$$\begin{cases} (\mathbf{v}_X^i)^\star = \left[(\mathbf{K}_X^i)^T\mathbf{E}_X\mathbf{K}_X^i + n^i\lambda^i\mathbf{K}_X^i\right]^{-1}(\mathbf{K}_X^i)^T\mathbf{E}_X\mathbf{c}^i \\ (b_X^i)^\star = \frac{\mathbf{1}}{n^i}\left((\mathbf{1})^T\mathbf{c}^i - (\mathbf{1})^T\mathbf{K}_X^i\left(\mathbf{v}_X^i\right)^\star\right) \end{cases} \tag{20}$$

where $\mathbf{E}_X = I - \frac{1}{n^i}\mathbf{1}\mathbf{1}^T$, and I is identity matrix with proper dimension. Substituting $(\mathbf{v}_X^i)^\star$ and $(b_X^i)^\star$ into (18), we can get

$$\begin{aligned} J_X^L(i) &= \frac{1}{n^i}(\mathbf{c}^i)^T\Big[\mathbf{E}_X - \mathbf{E}_X\mathbf{K}_X^i\left((\mathbf{K}_X^i)^T\mathbf{E}_X\mathbf{K}_X^i + \right. \\ &\qquad \left. n^i\lambda_X^i\mathbf{K}_X^i\right)^{-1}(\mathbf{K}_X^i)^T\mathbf{E}_X\Big]\mathbf{c}^i \\ &= (\mathbf{c}^i)^T\mathbf{L}_X^i\mathbf{c}^i \\ &= \mathbf{c}^T(S^i)^T\mathbf{L}_X^iS^i\mathbf{c} \end{aligned}$$

where

$$\begin{aligned}\mathbf{L}_X^i = \frac{1}{n^i}\Big[\mathbf{E}_X - \mathbf{E}_X\mathbf{K}_X^i((\mathbf{K}_X^i)^T\mathbf{E}_X\mathbf{K}_X^i + n^i\lambda_X^i\mathbf{K}_X^i)^{-1} \\ (\mathbf{K}_X^i)^T\mathbf{E}_X\Big].\end{aligned}$$

Also, we rewrite $\mathbf{L}_X^i$ as

$$\mathbf{L}_X^i = \lambda_X^i\mathbf{E}_X\left(n^i\lambda_X^i\mathbf{K}_X^i + \mathbf{E}_X\mathbf{K}_X^i(\mathbf{K}_X^i)^T\mathbf{E}_X\right)^{-1}\mathbf{E}_X. \tag{21}$$

Similarly, we have the formulation of J_X^L is

$$\begin{aligned} J_X^L &= \sum_{i=1}^n J_X^L(i) = \sum_{i=1}^n (\mathbf{c}^i)^T\mathbf{L}_X^i\mathbf{c}^i \\ &= \sum_{i=1}^n \left[\mathbf{c}^T(\mathbf{S}^i)^T\mathbf{L}_X^i\mathbf{S}^i\mathbf{c}\right] \\ &= \mathbf{c}^T\sum_{i=1}^n \left[(\mathbf{S}^i)^T\mathbf{L}_X^i\mathbf{S}^i\right]\mathbf{c}. \end{aligned} \tag{22}$$

Let $\mathbf{L}_X = \sum_{i=1}^n \left[(\mathbf{S}^i)^T\mathbf{L}_X^i\mathbf{S}^i\right]$, then

$$J_X^L = \mathbf{c}^T\mathbf{L}_X\mathbf{c}. \tag{23}$$

To summarize the objective function on local consistency in word space, we unified the formulations with two forms of predictors into (24)

$$J_X^L = \mathbf{c}^T\mathbf{L}_X\mathbf{c} \tag{24}$$

where

$$L_X = \sum_{i=1}^n \left[(\mathbf{S}^i)^T\mathbf{L}_X^i\mathbf{S}^i\right]$$

and

$$L_X^i = \begin{cases} \lambda_X^i\mathbf{E}_X\left(n^i\lambda_X^i\mathbf{I} + \mathbf{E}_X\mathbf{N}^i(\mathbf{N}^i)^T\mathbf{E}_X\right)^{-1}\mathbf{E}_X \\ \qquad\qquad \text{linear} \\ \lambda_X^i\mathbf{E}_X\left(n^i\lambda_X^i\mathbf{K}_X^i + \mathbf{E}_X\mathbf{K}_X^i(\mathbf{K}_X^i)^T\mathbf{E}_X\right)^{-1}\mathbf{E}_X \\ \qquad\qquad \text{kernel}. \end{cases} \tag{25}$$

2) Global Consistency in Word Space: Besides the local regularization in word space, we also require that cluster assignment should be smooth over the whole space. Let $\mathbf{W}_X \in \Re^{n\times n}$ be the sysmmetric weight matrix, in which each element $\mathbf{W}_X(i,j)$ measures the similarity between the i-th and the j-th words, and is assumed to be nonnegative. Similarly to [28], we have the following objective function:

$$J_X^G = \sum_{i\neq j}(c_i - c_j)^2 \mathbf{W}_X(i,j). \tag{26}$$

Intuitively, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar, that is $\mathbf{W}_X(i,j)$ is larger, $|c_i - c_j|$ should be smaller.

Define the Laplacian matrices $\mathbf{G}_X$ and its corresponding diagonal matrix $\mathbf{D}_X$ as

$$\mathbf{G}_X = \mathbf{D}_X - \mathbf{W}_X \tag{27}$$

where $(\mathbf{D}_X)_{ii} = \sum_{i\neq j}\mathbf{S}_X(i,j)$ for $\forall i$. Then global consistency loss in word space J_X^G can be deducted to

$$J_X^G = \mathbf{c}^T\mathbf{G}_X\mathbf{c}. \tag{28}$$

C. Modeling the Interdocument Relationship

To reveal interdocument relationship, we also need to model it by considering both local and global consistency in document space.

1) Local Consistency in Document Space: Similarly to the local regularization in word space, we choose linear regression and kernel regression in document space. Define f_Y^i be a local label predictor. Equation (29) fits predict f_Y^i with linear regression, and (30) is with kernel regression

$$f_Y^i(\mathbf{y}_j) = \mathbf{y}_j^T\mathbf{w}_Y^i + b_Y^i \tag{29}$$

$$f_Y^i(\mathbf{y}_j) = \psi(\mathbf{y}_j)\tilde{\mathbf{w}}_Y^i + b_Y^i \tag{30}$$

where $\mathbf{y}_j \in \mathcal{M}^i$, $\mathbf{w}_Y^i \in \Re^{n\times 1}$, $b_Y^i \in \Re$ and $\psi : \Re^n \to \mathcal{F}_Y$ be a mapping to map the document $\mathbf{y}$ into a high dimensional space F_Y.

Let $\mathbf{E}_Y = \mathbf{I} - \frac{1}{m^i}\mathbf{1}\mathbf{1}^T$, $\mathbf{1}$ be a m^i-dimensional column vector of all ones, also we have $\mathbf{d}^i = \mathbf{R}^i\mathbf{d}$, where $\mathbf{R}^i = [\mathbf{e}_k]^T, k \in \mathcal{B}^i$ is a selection matrix for each subset $\mathcal{M}^i$, $\mathbf{e}_k$ is a m-dimensional vector whose k-th element is one and all others are zeros. Define $\mathbf{K}_Y^i \in \Re^{m^i\times m^i}$ be the kernel matrix for $\mathbf{M}^i$ with its (i,j)-th entry $\mathbf{K}_Y^i(k,j) = \mathbf{K}_Y(\mathbf{y}_{i_k}, \mathbf{y}_{i_j})$, where $\mathbf{y}_{i_j}$ is the j-th element of $\mathcal{M}^i$. Following the similar derivations in Section IV-B1, the local consistency loss in document space J_Y^L is:

$$J_Y^L = \mathbf{d}^T\mathbf{L}_Y\mathbf{d} \tag{31}$$

where

$$\mathbf{L}_Y = \sum_{i=1}^{m}\left[(\mathbf{R}^i)^T\mathbf{L}_Y^i\mathbf{R}^i\right]$$

and

$$\mathbf{L}_Y^i = \begin{cases} \lambda_Y^i\mathbf{E}_Y\left(m^i\lambda_Y^i\mathbf{I} + \mathbf{E}_Y\mathbf{M}^i(\mathbf{M}^i)^T\mathbf{E}_Y\right)^{-1}\mathbf{E}_Y & \text{linear} \\ \lambda_Y^i\mathbf{E}_Y\left(m^i\lambda_Y^i\mathbf{K}_Y^i + \mathbf{E}_Y\mathbf{K}_Y^i(\mathbf{K}_Y^i)^T\mathbf{E}_Y\right)^{-1}\mathbf{E}_Y & \text{kernel.} \end{cases} \tag{32}$$

2) Global Consistency in Document Space: Let $\mathbf{W}_Y \in \Re^{m\times m}$ be the sysmmetric weight matrix with the element $\mathbf{W}_Y(i,j)$ as the similarity between the i-th and the j-th documents. Similarly to Section IV-B2, we have

$$J_Y^G = \sum_{i\neq j}(d_i - d_j)^2\mathbf{W}_Y(i,j). \tag{33}$$

Define the Laplacian matrices $\mathbf{G}_Y$ and its corresponding diagonal matrix $\mathbf{D}_Y$ as

$$\mathbf{G}_Y = \mathbf{D}_Y - \mathbf{W}_Y \tag{34}$$

where $(\mathbf{D}_Y)_{ii} = \sum_{i\neq j}\mathbf{S}_Y(i,j)$ for $\forall i$. Then the global consistency loss in document space J_Y^G can be deducted to

$$J_Y^G = \mathbf{d}^T\mathbf{G}_Y\mathbf{d}. \tag{35}$$

D. Unified Problem Formulation for CCLGC

By integrating all the above fitting errors from three-folds, that is, J_{XY} in (4) and (5) from word-document relationship, J_X^L in (24) and J_X^G in (28) from interword relationship, J_Y^L in (31) and J_Y^G in (35) from interdocument relationships, we have the overall loss function of our problem as follows:

$$\begin{aligned} J &= J_{XY} + \alpha(J_X^L + J_Y^L) + \beta(J_X^G + J_Y^G) \\ &= J_{XY} + \alpha[\mathbf{c}^T\mathbf{L}_X\mathbf{c} + \mathbf{d}^T\mathbf{L}_Y\mathbf{d}] + \beta[\mathbf{c}^T\mathbf{G}_X\mathbf{c} + \mathbf{d}^T\mathbf{G}_Y\mathbf{d}] \\ &= J_{XY} + \mathbf{c}^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c} + \mathbf{d}^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d} \end{aligned} \tag{36}$$

the parameter α, β in (36) are the weights to balance these terms. Then, our objective becomes to solve the following optimization problem:

$$\arg\min_{\mathbf{c},\mathbf{d}} J_{XY} + \mathbf{c}^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c} + \mathbf{d}^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}. \tag{37}$$

Definition 1: An optimal co-clustering is to solve (37) subject to the constraint on the number of word and document clusters.

V. Iterative Optimizing Rules and Algorithm

Note that the loss function on word-document relationship can be expressed by two forms, that is, (4) and (5). Then, from (4), the overall loss function J can be rewritten as

$$\begin{aligned} J = \sum_{\hat{\mathcal{X}}}\sum_{\mathbf{x}\in\hat{\mathcal{X}}} & p(\mathbf{x})D\left(p(Y|\mathbf{x})||q(Y|\hat{\mathcal{X}})\right) \\ & + \mathbf{c}^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c} \\ & + \mathbf{d}^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}]. \end{aligned} \tag{41}$$

And from (5), the overall loss function J can be rewritten as

$$\begin{aligned} J = \sum_{\hat{\mathcal{Y}}}\sum_{\mathbf{y}\in\hat{\mathcal{Y}}} & p(\mathbf{y})D\left(p(X|\mathbf{y})||q(X|\hat{\mathcal{Y}})\right) \\ & + \mathbf{d}^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d} \\ & + \mathbf{c}^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c}. \end{aligned} \tag{42}$$

Similarly to ITCC, we iteratively optimize the objective function (37) by word clustering (41) and document clustering (42). Furthermore, it can be proved that this interactive process

Algorithm 1 Co-Clustering via Local and Global Consistency (CCLGC)

Input:

Data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] = \mathbf{Y} = [\mathbf{y}_1^T; \cdots; \mathbf{y}_m^T]$
Number of word clusters: $\hat{n}$
Number of document clusters $\hat{m}$
Size of neighborhood k
Local regularization parameters on word space $\{\lambda_X^i\}_{i=1}^n$
Local regularization parameters on document space $\{\lambda_Y^i\}_{i=1}^m$
Objective function parameters α, β

Output:

Cluster assignment for word set $\mathbf{c}$
Document cluster assignment $\mathbf{d}$

1: **Initialization:**
Set $t = 0$, $p(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{X}}{\mathbf{1}^T\mathbf{X}\mathbf{1}}$
Initialize word and document cluster set $\mathbf{c}^{(0)}$, $\mathbf{d}^{(0)}$;
Calculate L_X, L_Y, G_X, G_Y by using (25) (32) (27) (34);
Compute $q^{(0)}(\hat{X}, \hat{Y})$, $q^{(0)}(X|\hat{X})$, $q^{(0)}(Y|\hat{Y})$, and the distribution $q^{(0)}(Y|\hat{\mathcal{X}})$

2: Word clustering: For each word $\mathbf{x}_i$, find its new cluster index c_i^{t+1} as

$$c_i^{(t+1)} = \arg\min_{c_i} D(p(Y|\mathbf{x}_i) \| q(Y|\hat{\mathcal{X}}^{(t)})) + (\mathbf{c}^{(t)})^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c}^{(t)}. \quad (38)$$

Let $\mathbf{d}^{(t+1)} = \mathbf{d}^{(t)}$.

3: Compute distribution: $q^{(t+1)}(\hat{X}, \hat{Y})$, $q^{(t+1)}(X|\hat{X})$, $q^{(t+1)}(Y|\hat{Y})$, and $q^{(t+1)}(X|\hat{\mathcal{Y}})$.

4: Document clustering: For each document $\mathbf{y}_j$, find its new cluster index as

$$d_j^{(t+2)} = \arg\min_{d_j} D(p(X|\mathbf{y}_j) \| q(X|\hat{\mathcal{Y}}^{(t+1)})) + \mathbf{d}^{(t+1)^T}(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}^{(t+1)}]. \quad (39)$$

Let $\mathbf{c}^{(t+2)} = \mathbf{c}^{(t+1)}$.

5: Compute distribution: $q^{(t+2)}(\hat{X}, \hat{Y})$, $q^{(t+2)}(X|\hat{X})$, $q^{(t+2)}(Y|\hat{Y})$, $q^{(t+2)}(Y|\hat{\mathcal{X}})$.

6: Compute objective function value using (41), and compute the change in objective function, i.e.,

$$\Delta^{(t)} = J^{(t)} - J^{(t+2)}. \quad (40)$$

If $\Delta^{(t)} < 10^{-3}$, Stop and return $\mathbf{c} = \mathbf{c}^{(t+2)}$, $\mathbf{d} = \mathbf{d}^{(t+2)}$
Else, set $t = t + 2$, and go to step 2.

can monotonically decreases the objective function, detailed in Theorem 1. With this intuition, we present the algorithm of CCLGC in Algorithm 1. Specifically, the algorithm starts with an initial co-clustering $\mathbf{c}^0$ and $\mathbf{d}^0$ (see step 1). Then, we fix the document clusters $\mathbf{d}$ and minimize the function in (41). Note that, when fix $\mathbf{d}$, $\mathbf{d}^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}$ is fixed, so minimize (41) is equal to minimize (43). We reassign each word $\mathbf{x}$ to a new word cluster to minimize (43) (see step 2). In third step, the algorithm computes the required marginals of $q(\cdot)$. The following document clustering, and minimize (42) is equal to minimize (44) (see step 4 and 5). This iterative process stops when the objective function no longer decreases

$$J = \sum_{\hat{\mathcal{X}}} \sum_{\mathbf{x}\in\hat{\mathcal{X}}} p(\mathbf{x}) D\left(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}})\right) + \mathbf{c}^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c} \quad (43)$$

$$J = \sum_{\hat{\mathcal{Y}}} \sum_{\mathbf{y}\in\hat{\mathcal{Y}}} p(\mathbf{y}) D\left(p(X|\mathbf{y}) || q(X|\hat{\mathcal{Y}})\right) + \mathbf{d}^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}. \quad (44)$$

Theorem 1: CCLGC Algorithm monotonically decreases the objective function given in (41) and (42).

Proof: Start from the t-th iteration, the objective function begins with $J^{(t)}$

$$J^{(t)} = \sum_{\hat{\mathcal{X}}^{(t)}} \sum_{\mathbf{x}\in\hat{\mathcal{X}}^{(t)}} p(\mathbf{x}) D\left(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}}^{(t)})\right) + (\mathbf{c}^{(t)})^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c}^{(t)} + (\mathbf{d}^{(t)})^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}^{(t)}.$$

Followed by step 2

$$\begin{aligned} &\sum_{\hat{\mathcal{X}}^{(t)}} \sum_{\mathbf{x}\in\hat{\mathcal{X}}^{(t)}} p(\mathbf{x}) D\left(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}}^{(t)})\right) \\ &\quad + (\mathbf{c}^{(t)})^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)(\mathbf{c}^{(t)}) \\ &\quad + (\mathbf{d}^{(t)})^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}^{(t)}] \\ \geq &\sum_{\hat{\mathcal{X}}^{(t)}} \sum_{\mathbf{x}\in\hat{\mathcal{X}}^{(t)}} p(\mathbf{x}) D\left(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}}_{c_x^{(t+1)}})\right) \\ &\quad + (\mathbf{c}^{(t+1)})^T(\alpha\ \mathbf{L}_X + \beta\mathbf{G}_X)\mathbf{c}^{(t+1)}] \\ &\quad + (\mathbf{d}^{(t)})^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}^{(t)}]. \end{aligned} \quad (45)$$

Proved by [2, Th. 4.1], we have

$$\begin{aligned} &\sum_{\hat{\mathcal{X}}^{(t)}} \sum_{\mathbf{x}\in\hat{\mathcal{X}}^{(t)}} p(\mathbf{x}) D\left(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}}^{(t)})\right) \\ \geq &\sum_{\hat{\mathcal{X}}^{(t+1)}} \sum_{\mathbf{x}\in\hat{\mathcal{X}}^{(t+1)}} p(\mathbf{x}) D\left(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}}^{(t+1)})\right) \end{aligned} \quad (46)$$

and

$$\mathbf{d}^{t+1} = \mathbf{d}^t. \quad (47)$$

Therefore

$$\begin{aligned} J^t \geq &\sum_{\hat{\mathcal{X}}^{(t)}} \sum_{\mathbf{x}\in\hat{\mathcal{X}}^{(t)}} p(\mathbf{x}) D\left(p(Y|\mathbf{x}) || q(Y|\hat{\mathcal{X}}^{(t)})\right) \\ &+ (\mathbf{c}^{(t+1)})^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)(\mathbf{c}^{(t+1)}) \\ &+ (\mathbf{d}^{(t+1)})^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}^{(t+1)}] \\ = &J^{t+1}. \end{aligned} \quad (48)$$

Similarly, following step 4, we have:

$$\begin{aligned} J^{t+1} = &\sum_{\hat{\mathcal{Y}}^{(t+1)}} \sum_{\mathbf{y}\in\hat{\mathcal{Y}}^{(t+1)}} p(\mathbf{y}) D\left(p(X|\mathbf{y}) || q(X|\hat{\mathcal{Y}}^{(t+1)})\right) \\ &+ (\mathbf{c}^{(t+1)})^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)(\mathbf{c}^{(t+1)}) \\ &+ (\mathbf{d}^{(t+1)})^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}^{(t+1)} \\ \geq &\sum_{\hat{\mathcal{Y}}^{(t+2)}} \sum_{\mathbf{y}\in\hat{\mathcal{Y}}^{(t+2)}} p(\mathbf{y}) D(p(X|\mathbf{y}) || q(X|\hat{\mathcal{X}}^{(t+2)})) \\ &+ (\mathbf{c}^{(t+2)})^T(\alpha\mathbf{L}_X + \beta\mathbf{G}_X)(\mathbf{c}^{(t+2)})] \\ &+ (\mathbf{d}^{(t+2)})^T(\alpha\mathbf{L}_Y + \beta\mathbf{G}_Y)\mathbf{d}^{(t+2)}] \\ = &J^{t+2}. \end{aligned} \quad (49)$$

By combining (48) and (49), it follows that the objective function never increases at every iteration. ■

We now analyze the computational complexity of the proposed CCLGC algorithm. In the initialization step, finding k nearest neighbors of each word and document takes $O(m^2n + m^2 \log m + n^2m + n^2 \log n)$, calculating L_X and L_Y according to (25) and (32) takes $O(m(nk^2+k^3))+n(mk^2+k^3)$, and calculating G_X and G_Y according to (27) and (34) takes $O(m+n)$. In each iteration, it takes $O(\varsigma(\hat{m}+\hat{n}+\hat{m}m^2+\hat{n}n^2))$, where ς is the number of nonzeros in $p(\mathbf{x}, \mathbf{y})$.

Remark 1: The time complexity of Algorithm 1 is $O((m+n)mn + m^2 \log m + n^2 \log n + (mn + mk + nk)k^2 + \varsigma(\hat{m}+\hat{n}+\hat{m}m^2+\hat{n}n^2)\tau)$, where τ is the number of iterations.

In the experiments, it is shown that $T = 20$ is enough for convergence.

The main memory cost of CCLGC is to store the pairwise distance matrices. Thus, the memory complexity is $O(m^2+n^2)$.

VI. Experiment

We conduct experiments on benchmark document datasets to empirically compare the clustering results of CCLGC with other clustering and co-clustering algorithms.

A. Datasets

We perform our experiments on two text datasets which are widely used in the experiments of clustering: 20 Newsgroup,[1] and WebKB.[2]

1) *20 Newsgroup:* This dataset consists of 20 000 messages which are taken from 30 Usenet newsgroup and separated into 20 different news topics. Similarly to [29], [14], we choose 2000 words with the largest contribution to the mutual information between the words and documents, and then discard the empty documents.
2) *WebKB:* This dataset contains WWW pages collected from computer science departments of four universities (Cornell, Texas, Washington, and Wisconsin) in January 1997. The 8282 pages were manually classified into the following categories, student, faculty, staff, department, course, project, and other. We select 1000 word with the largest contribution to the mutual information between the words and documents for this dataset.

B. Evaluation Metrics and Baselines

It is a nontrivial task to validate the clustering and co-clustering results. In the evaluation, we choose two metrics, accuracy (AC), and normalized mutual information (NMI), to measure the clustering performance.

1) *Accuracy (AC):* The first performance measure we choose is the AC, which discovers the relationship between obtained cluster label and the groundtruth on each data. For a document sample $\mathbf{y}_i$, the cluster label d_i is estimated by different clustering or co-clustering algorithm, and the groundtruth cluster label is defined as l_i. The AC is defined as

$$AC = \frac{\sum_{i=1}^{m} \delta(l_i, map(q_i))}{m} \tag{50}$$

where $\delta(x, y)$ is the delta function that equals one if $x = y$, and equals zero other wise. $map(q_i)$ is the permutation mapping function that map each cluster label q_i to the equivalent label from dataset. The mapping process can be refer to Kuhn–Munkres algorithm [30].
2) *Normalized Mutual Information (NMI):* The NMI is widely used to determine the quality of clusters. For two random variables X and Y, it is defined as [31]

$$NMI = \frac{I(X, Y)}{H(X)H(Y)} \tag{51}$$

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. Given a clustering results, the NMI in (51) is estimated as [31]

$$NMI = \frac{\sum_{i=1}^{\hat{m}} \sum_{j=1}^{\hat{m}} \hat{m}_{ij} \log\left(\frac{m\hat{m}_{ij}}{\hat{m}_i(\hat{m}^\star)_j}\right)}{\sqrt{\left(\sum_{i=1}^{\hat{m}} \hat{m}_i \log\frac{\hat{m}_i}{m}\right)\left(\sum_{j=1}^{\hat{m}} (\hat{m}^\star)_j \log\frac{(\hat{m}^\star)_j}{m}\right)}} \tag{52}$$

where $\hat{m}_i$ is the number of data contained in the i-th obtained cluster, $(\hat{m}^\star)_j$ is the number of data in the j-th groundtruth class, and $\hat{m}_{ij}$ denotes the number of data that are in the intersection between the i-th obtained cluster and the j-th groundtruth class. Obviously, the large this value, the better the performance of clustering.

We compare our method with the following six baselines.

1) Kmeans [32].
2) Normalized cut (N-cut) [33].
3) Clustering with local and global regularization (LCLGR denotes the approach using regularized linear classifiers as its local predictors, and KCLGR represents the kernel one) [24].
4) ITCC [2].
5) Locally discriminative co-clustering (LDCC) [14].
6) ITCC incorporating word similarities and document similarities (ITCCWDS) [13].

The first three methods are all one-side clustering algorithms, while the following three are co-clustering ones. LCLGR and KCLGR consider both local and global consistency. In co-clustering algorithms, ITCC only focuses on the relationship between word and document but neglects the within relationships, LDCC considers the local patch in the word and document spaces, but discards the global information from these two spaces, while ITCCWDS only takes the global word/document similarities into consideration.

For our proposed co-clustering method, we use LCCLGC to represent the approach with local linear regularization, and use KCCLGC to indicate that with local kernel regularization.

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/.
[2] http://www.cs.cmu.edu/ WebKB/.

C. Experiment Setup

1) Parameter Settings: In the experiments, the reported performances are averaged over 10 independent runs. For parameter settings, we ran each algorithm by using different parameters, and reported the best result. For one-side clustering algorithms, the number of document clusters is set equal to the groundtruth. For co-clustering algorithms, the numbers of word and document clusters are both set to the true number of classes for all the datasets.

In K-means, the cluster centers are randomly initialized.

In N-cut, the similarity between two samples $\mathbf{y}_i$ and $\mathbf{y}_j$ is computed with

$$\mathbf{W}(i,j) = \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|_2^2}{\sigma}\right) \tag{53}$$

where σ is searched from the grid $\{1e-2, 1e-1, 1, 1e1, 1e2\}$, which is set to be the same as that in [14].

In LCLGR and KCLGR, the parameter settings are the same as those in [24]. Specifically, the neighborhood size is searched from the grid: $\{5, 10, 20, 50, 100, 150, 200\}$, and the parameters λ_i of the local regularized classifier are set to be the same and searched from the grid $\{0.01, 0.1, 1, 10\,100\}$. The pairwise data similarities used for constructing the global smoothness regularizer are calculated using Gaussian kernel. The balanced parameter α is searched from the grid $\{1e-3, 1e-2, 1e-1, 1, 1e1, 1e2\}$. For KCLGR, the Gaussian kernel is adopted for constructing the local kernel predictors with its width tuned from the grid $\{4e-3\sigma_0, 4e-2\sigma_0, 4e-1\sigma_0, \sigma_0, 4e1\sigma_0, 4e2\sigma_0, 4e3\sigma_0\}$, where σ_0 is the mean distance between any two examples in the data set.

In LDCC, the parameter λ for local linear regression, the neighborhood size, and balanced parameters α, β are set to be the same as those in [14]. The setting of neighbor size k is the same as those in LCLGR and KCLGR.

In ITCCWDS, the balanced parameters α and β are set to be the same as those in LDCC.

In our proposed method, the parameter λ_X^i and λ_Y^i are set to be the same, and their grid ranges are set to be the same as λ_i in LCLGR and KCLGR. The balanced parameters α and β are set to be the same as LDCC. Since our method converges to a local minimum, which does not guarantee the global one, the choices of $\mathbf{c}^0$ and $\mathbf{d}^0$ in step 1 is important. We adopt different strategies for these two initializations. For word cluster initialization, we choose initial word cluster "centroids" to be maximally far apart from each other. First, we take the word which is the farthest to the whole data set as the first word cluster "centroid." Then, take the word which is the farthest to all the confirmed "centroid," until we get $\hat{m}$ "centroids." For document initialization, we first compute the "mean" vector of all the documents, then randomly perturb this vector to get $\hat{n}$ starting document clusters.

D. Sensitivity to Parameter k, α, and β

In this subsection, we investigate the sensitivity with respect to the neighborhood size k, and the balanced parameters α, β. When we adjust the values of k, α, and β, respectively, we keep the other parameters fixed at the optimal values. The performance comparison under different settings is evaluated on both the 20 Newsgroup and the unbalanced WebKB dataset. We select the number of document classes *num* as two in the 20 Newsgroup, and as four in the WebKB. Other experimental settings are the same as these in Section VI-C1. We plot the clustering AC with respect to k, α, and β for the 20 Newsgroup dataset in Fig. 3(a)–(c), and that for WebKB dataset in Fig. 4(a)–(c).

As we can see, CCLGC is a little sensitive to the neighborhood size of the graph. Fortunately, it usually achieves good result when the neighborhood size is not too large, e.g., $k \in [10, 20]$ in our experiments. On the other hand, a proper trade-off between local and global consistency can bring a better result in LCCLGC and KCCLGC. The balanced parameters α and β are suggested to be set between 0.1 and 10. The selection of α and β also experimentally support the reasonability of considering local and global consistency.

E. Experiment Results

1) Performance on Balanced Datasets: This section shows the performance comparison on balanced datasets, that is, all the classes are of the same size.

In 20 Newsgroup, we conducted experiment with different number of document classes *num*, which ranges from 2 to 10. The selected document classes *num* are chosen randomly from whole dataset. For each document class, we randomly select 300 documents. We conducted 10 runs for each selected class number *num*, and report the average AC and NMI in Tables II and III.

Considering that the class sizes in WebKB are not equal to each other, we choose subset of this dataset to generate the balanced classes. We conducted experiment with $num(= 2, \ldots, 7)$ document classes, and randomly selected 100 documents for each selected class. Also, 10 runs for each selected class number *num* are conducted, and the average AC and NMI are reported in Tables IV and V.

From these results, we have the following observations.

1) In most cases, our proposed CCLGR achieves the best performance, and KCCLGC usually outperforms LCCLGC since the local distributions of the dataset are nonlinear. We noted that the results of our method are not satisfied on balanced WebKB with class number as 2, this is mainly because that the global consistency is hard to reveal on the small test set, which is only 100 documents in each cluster and 200 documents totally in this experiment. Ncut and KCLGR yield good results on this test set as they are based on local regularizer, which is more suitable when dataset is small.
2) The results of Kmeans are mostly poorer than other methods, since the distributions of the text data are commonly complicated than mixtures of spherical Gaussians.
3) The mixed-regularization algorithms, including LCLGR, KCLGR, and LCCLGC, KCCLGC mostly perform better than the algorithms based on a single regularizer, since they make use of more information from the dataset.
4) The incorporation of interword and interdocument relationships increases the co-clustering performance.

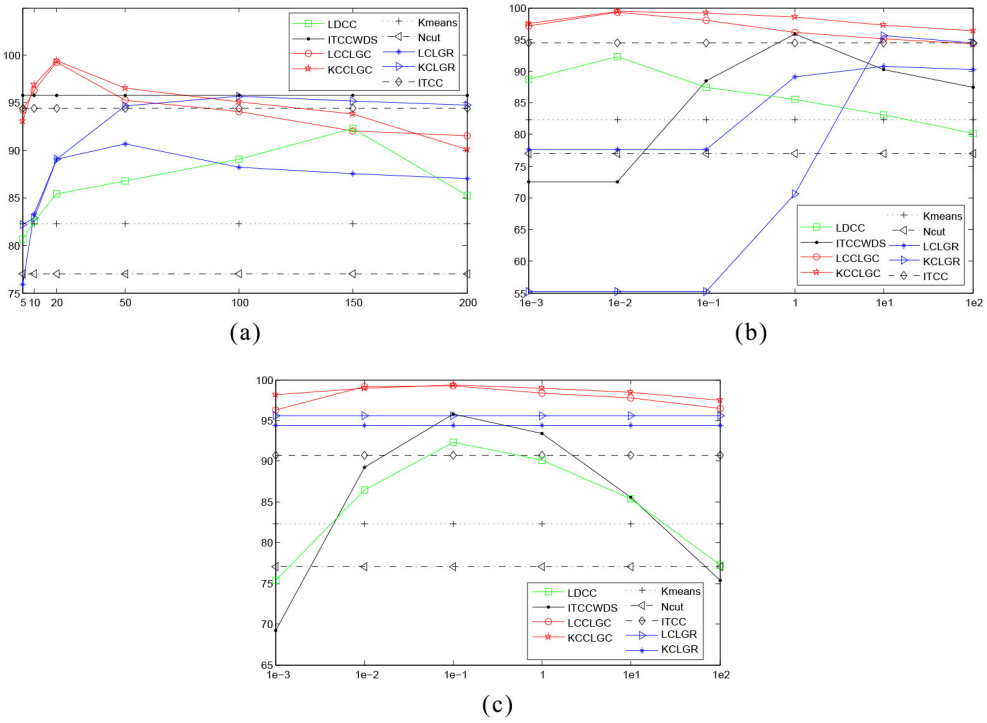

Fig. 3. Parameter sensitivity analysis on the 20 Newsgroup dataset(%). (a) Performance comparison with varying number of neighbors. (b) Performance comparison against different α. (c) Performance comparison against different β.

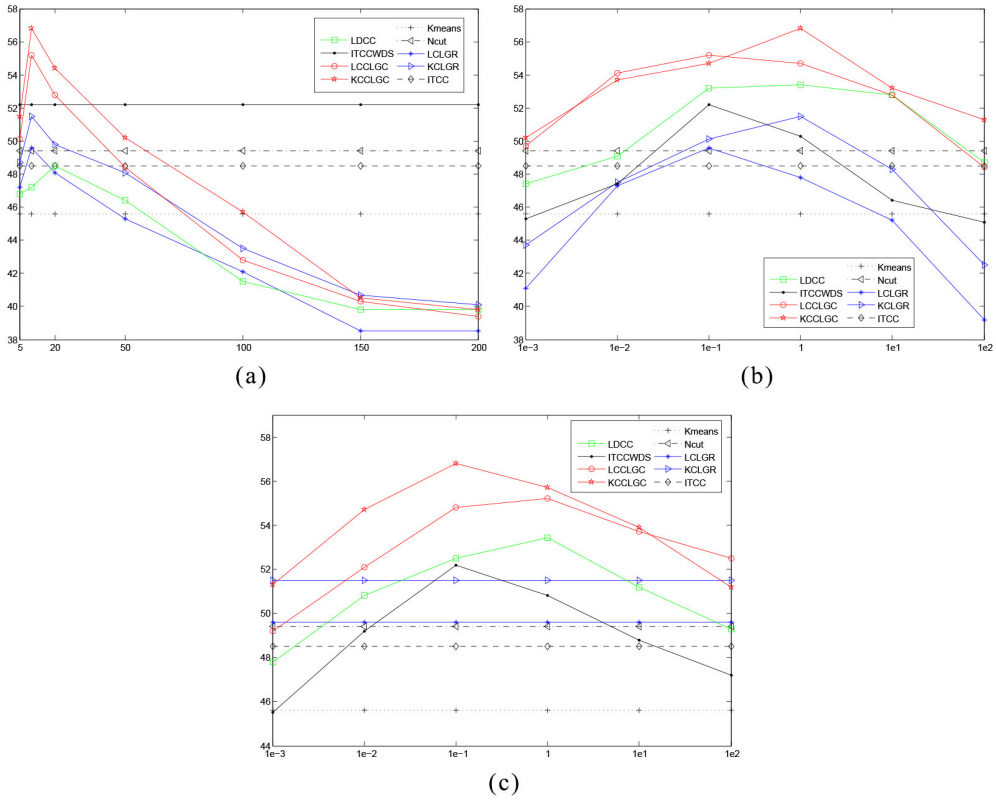

Fig. 4. Parameter sensitivity analysis on the WebKB dataset(%). (a) Performance comparison with varying number of neighbors. (b) Performance comparison against different α. (c) Performance comparison against different β.

5) As we expected, LCCGC and KCCLGC generally outperform LDCC and ITCCWDS, since the local predicted labels are smoothed by the global smoother.

2) Performance on Unbalanced Dataset: In order to comprehensively evaluate the proposed method, we also perform it on the unbalanced dataset, that is, the sizes of document

TABLE II
COMPARISON OF ACCURACY (%) ON 20 NEWSPAPER (MEAN±STD)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
2	82.3 ±8.4	77.0 ±23.2	90.7 ±2.4	95.6 ±2.4	94.4 ±2.5	92.3 ±3.4	95.8 ±5.8	**99.3** ±0.5	**99.4** ±0.3
3	81.2 ±6.2	66.8 ±21.7	82.5 ±15.6	82.5 ±15.7	79.9 ±12.8	84.4 ±10.2	82.2 ±9.8	**86.3** ±8.9	**88.1** ±9.4
4	67.32 ±11.93	62.07 ±15.56	79.4 ±10.1	79.6 ±9.8	72.1 ±11.1	82.1 ±8.4	78.6 ±5.5	**84.6** ±6.7	**85.7** ±6.9
5	64.3 ±10.4	62.5 ±16.5	69.9 ±9.4	72.5 ±10.5	65.6 ±12.9	**78.8** ±7.3	73.8 ±8.5	78.4 ±8.5	**80.5** ±7.2
6	60.1 ±8.1	57.4 ±8.8	66.6 ±7.3	67.4 ±7.6	62.6 ±9.1	70.9 ±7.8	70.1 ±6.9	**72.3** ±6.7	**75.2** ±7.0
7	56.1 ±5.6	55.5 ±8.2	64.2 ±6.0	64.0 ±6.9	61.8 ±5.4	**71.3** ±6.1	69.2 ±5.4	70.4 ±6.2	**73.4** ±6.8
8	53.9 ±9.3	52.4 ±5.9	60.8 ±10.5	62.5 ±10.7	54.8 ±9.3	64.4 ±9.5	64.5 ±7.3	**67.5** ±5.6	**69.3** ±5.1
9	54.0 ±7.5	50.6 ±8.5	63.4 ±8.8	63.4 ±9.6	54.3 ±8.4	66.7 ±7.2	65.0 ±5.2	**68.5** ±5.5	**69.2** ±6.06
10	48.8 ±5.0	49.0 ±3.3	56.8 ±5.2	58.7 ±5.3	51.3 ±4.8	63.1 ±5.4	60.4 ±5.0	**65.6** ±4.4	**67.8** ±5.1

TABLE III
COMPARISON OF NORMALIZED MUTUAL INFORMATION (%) ON 20 NEWSPAPER (MEAN±STD)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
2	56.2 ±20.0	47.8 ±18.5	65.2 ±10.0	76.0 ±10.8	70.8 ±10.4	72.9 ±13.4	75.6 ±17.1	**94.9** ±3.0	**94.3** ±2.9
3	54.0 ±10.4	42.3 ±8.4	60.4 ±17.0	60.8 ±17.8	53.1 ±15.6	62.7 ±10.2	60.0 ±13.2	**64.6** ±14.3	**70.3** ±13.2
4	44.2 ±9.3	41.5 ±9.5	58.6 ±13.4	59.0 ±13.0	49.4 ±12.4	58.1 ±11.5	57.3 ±8.0	**60.2** ±9.5	**64.8** ±12.4
5	42.6 ±7.2	40.7 ±8.6	53.4 ±8.3	54.5 ±9.8	47.4 ±8.7	56.7 ±9.6	58.2 ±10.7	**58.9** ±8.7	**62.6** ±8.6
6	42.1 ±7.3	35.2 ±7.1	54.4 ±7.7	54.9 ±7.8	46.6 ±7.7	56.8 ±8.4	57.4 ±8.5	**58.2** ±7.4	**60.7** ±7.5
7	41.0 ±5.8	34.8 ±7.3	52.8 ±5.0	52.6 ±5.6	46.2 ±5.4	54.6 ±7.5	54.3 ±8.0	**56.7** ±8.7	**58.5** ±6.7
8	39.6 ±5.4	32.2 ±6.4	48.2 ±8.0	49.4 ±7.5	44.0 ±6.7	50.3 ±7.3	49.8 ±6.8	**52.6** ±6.5	**55.0** ±6.2
9	38.4 ±3.5	33.3 ±4.9	51.7 ±5.6	51.6 ±6.6	44.9 ±5.3	50.6 ±6.8	49.7 ±7.0	**53.6** ±5.3	**54.5** ±5.7
10	37.1 ±4.8	32.4 ±3.8	47.5 ±4.0	48.1 ±4.1	43.9 ±4.2	50.4 ±5.0	48.3 ±5.7	**51.8** ±4.9	**52.7** ±4.7

TABLE IV
COMPARISON OF ACCURACY (%) ON BALANCED WEBKB (MEAN±STD)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
2	79.4 ±9.6	**83.4** ±8.7	83.2 ±9.4	**84.5** ±7.9	77.6 ±11.3	84.2 ±8.8	79.5 ±10.5	80.0 ±9.8	80.3 ±9.7
3	70.8 ±13.7	73.1 ±13.3	69.5 ±14.6	73.7 ±10.3	70.0 ±9.9	75.2 ±10.8	74.7 ±11.3	**76.2** ±9.6	**78.4** ±9.4
4	68.3 ±9.5	69.4 ±10.6	68.5 ±9.8	71.6 ±9.1	68.5 ±10.3	72.5 ±9.8	71.4 ±9.2	**73.5** ±9.3	**74.3** ±9.8
5	57.7 ±6.4	62.4 ±9.4	60.6 ±8.3	63.5 ±7.8	57.4 ±7.7	64.8 ±8.0	63.3 ±7.5	**65.4** ±7.0	**67.2** ±7.4
6	51.2 ±6.1	55.3 ±5.3	52.4 ±7.4	52.8 ±7.1	50.5 ±5.7	56.2 ±6.4	55.3 ±5.5	**58.2** ±6.4	**59.7** ±5.7
7	45.3 ±5.1	44.2 ±4.6	47.1 ±6.4	47.8 ±6.1	46.5 ±4.3	54.0 ±5.5	53.8 ±4.7	**55.2** ±4.4	**57.6** ±4.6

classes are largely skewed. This experiment is conducted on WebKB, in which the size of document class ranges from 137 to 3728. The number of document classes is selected in the experiment as $num = 2, \ldots, 7$. Ten runs are conducted for each *num*. In each run, we select 10% documents from each selected class, and generate the highly unbalanced subsets. The average results are reported in Tables VI and VII.

From these results, we have the following observations.

1) In this unbalanced dataset, our proposed CCLGR achieves the best performance, and KCCLGC usually outperforms LCCLGC since the local distributions of the dataset are nonlinear. Moreover, LCCLGC and KCCLGC perform better results on $num = 2$ than the pure-local-regularization method, includes Ncut, LDCC,

TABLE V
COMPARISON OF NORMALIZED MUTUAL INFORMATION (%) ON BALANCED WEBKB (MEAN±STD)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
2	42.3 ±20.8	**45.3** ±24.5	43.7 ±16.2	**47.0** ±20.2	40.6 ±18.1	43.5 ±19.6	42.0 ±18.4	42.7 ±17.4	42.6 ±18.5
3	39.2 ±15.4	39.5 ±13.5	40.0 ±14.7	42.7 ±13.4	38.2 ±17.5	42.2 ±15.8	41.4 ±17.5	**43.5** ±13.8	**44.1** ±14.1
4	40.8 ±9.3	41.7 ±9.2	42.6 ±9.6	42.9 ±9.7	41.1 ±10.4	43.0 ±9.2	42.8 ±9.6	**44.0** ±9.5	**45.2** ±9.4
5	33.4 ±6.6	34.9 ±8.1	37.1 ±7.3	38.2 ±7.5	36.7 ±8.4	38.5 ±7.2	37.6 ±6.3	**40.2** ±6.8	**41.8** ±6.4
6	32.4 ±5.2	34.3 ±6.4	35.9 ±6.3	36.1 ±6.2	34.1 ±5.4	36.7 ±5.5	34.9 ±5.1	**38.3** ±5.3	**38.8** ±5.9
7	31.1 ±2.5	28.7 ±1.0	32.4 ±2.1	34.5 ±2.6	33.4 ±2.1	37.6 ±1.4	35.4 ±2.2	**38.2** ±1.6	**38.8** ±1.1

TABLE VI
COMPARISON OF ACCURACY (%) ON UNBALANCED WEBKB (MEAN±STD)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
2	66.8 ±14.5	68.2 ±14.3	69.7 ±13.7	71.8 ±13.8	61.7 ±12.8	73.8 ±13.5	71.2 ±12.1	**74.2** ±12.5	**76.3** ±12.9
3	51.5 ±11.0	56.6 ±9.9	58.4 ±10.3	58.4 ±9.4	53.6 ±9.4	58.1 ±7.7	57.5 ±9.3	**60.3** ±10.7	**62.4** ±9.8
4	45.6 ±4.7	49.4 ±5.2	49.6 ±6.7	51.5 ±5.4	48.5 ±4.5	53.4 ±3.9	52.2 ±4.3	**55.2** ±4.6	**56.8** ±4.7
5	45.9 ±10.5	43.9 ±9.8	48.4 ±9.4	50.2 ±9.8	43.8 ±5.0	52.4 ±10.7	52.5 ±9.6	**54.0** ±8.4	**54.7** ±9.1
6	40.2 ±6.6	41.1 ±4.7	43.5 ±5.8	43.8 ±5.2	40.6 ±5.5	47.7 ±4.9	45.3 ±4.5	**48.2** ±4.4	**49.3** ±4.7
7	38.7 ±3.2	38.5 ±0.0	41.1 ±2.4	42.3 ±2.0	36.0 ±2.1	43.2 ±1.2	42.7 ±1.8	**44.5** ±1.1	**45.4** ±1.3

TABLE VII
COMPARISON OF NORMALIZED MUTUAL INFORMATION (%) ON UNBALANCED WEBKB (MEAN±STD)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
2	17.1 ±18.6	20.0 ±19.5	15.4 ±19.7	16.1 ±18.5	11.7 ±17.5	18.5 ±14.6	17.4 ±15.8	19.5 ±15.3	**20.5** ±14.7
3	16.3 ±12.8	19.2 ±11.6	18.5 ±11.9	19.1 ±12.2	16.9 ±8.7	20.0 ±8.9	19.5 ±9.2	**21.4** ±9.5	**23.3** ±9.4
4	16.1 ±4.4	18.5 ±4.5	16.6 ±4.7	20.4 ±5.1	19.5 ±4.4	23.0 ±3.4	21.7 ±3.8	**25.7** ±4.3	**25.5** ±4.9
5	22.7 ±12.2	19.4 ±8.9	20.9 ±9.4	22.6 ±9.7	20.7 ±6.1	25.4 ±9.4	23.2 ±8.5	**25.5** ±7.3	**26.8** ±8.2
6	20.1 ±6.4	19.4 ±4.5	19.9 ±5.9	21.2 ±6.3	20.7 ±4.9	24.6 ±4.3	24.5 ±4.7	**26.3** ±4.4	**26.6** ±4.7
7	19.5 ±0.6	19.1 ±0.0	20.4 ±1.7	21.8 ±1.1	19.7 ±1.3	23.2 ±1.5	22.1 ±1.0	**23.9** ±0.9	**24.5** ±1.2

since the test dataset includes more documents than balanced one and the global consistency is easier to recover.

2) In general, the AC keeps decreasing as the number of classes *num* increases. Bur NMI does not follow this pattern. This is mainly because NMI is quite complex and it depends on the specific dataset [14].
3) As we expected, the proposed algorithms LCCLGC and KCCLGC generally outperform LDCC and ITCCWDS, since they reveal global consistency on the dataset. On the other hand, they also outperform LCLGR and KCLGR, since they make use of the relationships between words and documents and simultaneously group them into word clusters and document clusters.

Since the unbalanced datasets are more difficult to be clustered, the performance of all the method decreases in this experiment. Nevertheless, our proposed CCLGR method still achieves best for almost every case.

3) Discussion on Scalability: As discussed in Section V, computational complexity of our proposed algorithm is $O(m^2 log(m))$, where m is the number of samples. In this part, we will compare the computational time, AC, and NMI under the different dataset scalabilities. Three kinds of datasets are selected in this experiment. The first is 20 Newsgroup with document class number as 10 and the number of documents in each class is 300. The second one is 20 Newsgroup with all the 18 828 documents from 20 classes. And the last one is New York Times dataset, which consists of 421 384 news from five categories, and was downloaded from New York Times. All the experiments are conducted on a server with 2.27GHz CPU and 64GB memory. The code is implemented on MATLAB platform. The results are shown in

TABLE VIII
COMPARISON OF COMPUTATIONAL TIME (s), ACCURACY (%) AND NMI (%) ON 20 NEWSGROUP SUBSET ($\sharp Samples = 6000$)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
Computational Time	21	49	54	68	72	75	80	87	95
Accuracy	48.8	49.0	56.8	58.7	51.3	63.1	60.4	65.6	67.8
NMI	37.1	32.4	47.5	48.1	43.9	50.4	48.3	51.8	52.7

TABLE IX
COMPARISON OF COMPUTATIONAL TIME (s), ACCURACY (%), AND NMI (%) ON THE WHOLE 20 NEWSGROUP DATASET ($\sharp Sample = 18.828$)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
Computational Time	103	4683	4852	5320	6258	7492	7633	8203	8845
Accuracy	40.8	58.3	66.4	68.7	73.5	78.4	76.5	83.6	86.8
NMI	25.4	34.8	40.8	42.4	60.5	65.2	66.3	72.9	75.3

TABLE X
COMPARISON OF COMPUTATIONAL TIME (h), ACCURACY (%), AND NMI (%) ON NEW YORK TIMES DATASET ($\sharp Sample = 421384$)

	Kmeans	Ncut	LCLGR	KCLGR	ITCC	LDCC	ITCCWDS	LCCLGC	KCCLGC
Computational Time	1.3[1]	9.4	10.8	11.1	11.2	11.1	12.2	12.4	12.8
Accuracy	53.85	55.24	62.41	63.80	63.53	68.94	69.68	70.32	77.57
NMI	32.94	38.48	38.72	39.33	42.44	42.45	43.12	53.38	55.78

[1] The computational time of Kmeans was recorded under single thread, while that of others was recorded as parallel time under 5 threads.

Tables VIII, IX, and X, respectively. Considering the high time complexity of N-Cut, LCLGR, KCLGR, ITCC, LDCC, ITCCWDS, and our proposed LCCLGC, KCCLGC on the last dataset, we utilize multithreaded programming to obtain the results. All codes of those methods were split into five threads, and the computational time was recorded as parallel time.

From the results, we can see that, when dataset is small, our algorithms, LCCLGC and KCCLGC, take 12 and 20 more seconds than the 2nd competitor (LDCC), but achieve 2.5% and 5.7% higher on AC, respectively, shown in Table VIII. When dataset scalability increases to 18 828, the improvements of AC for LCCLGC and KCCLGC to the 2nd competitor (LDCC) are 5.2% and 8.4%, while the computational time is 80 times of Kmeans and 1200 seconds than the 2nd competitor (LDCC), shown in Table IX. When dataset scalability increases to 421 384, the computational time of all the methods except Kmeans is very huge. The improvements of AC for LCCLGC and KCCLGC to the 2nd competitor (ITCCWDS) are 7.2% and 10.6%, respectively, shown in Table X.

The reason of slight improvement on AC in small scale dataset might be because that as the small amount of samples hinder to explore the global consistency on samples. And we can see that, when the dataset scalability increases, the improvements on both AC and NMI increase too. The computational cost of our method is dramatically high when dealing with large scale dataset, as we need to search k nearest neighbors of each sample, whose time complexity is $O(m^2 log(m))$. From this observation, the scalability of our proposed method is not good enough, thus, we state this as an opportunity for our future work.

VII. CONCLUSION

In this paper, we have proposed a novel co-clustering method named CCLGC, which fully explores word-document, interword and interdocument relationships simultaneously. In modeling the interword and interdocument relationships, the proposed method preserves the merit of local and global learning methods. The experimental results show that CCLGC performs better than many state-of-art clustering and co-clustering methods.

In the future, considering that the real-world dataset usually mixes with some noises and its large scale characteristic, we will focus on the co-clustering method with those out-of-sample and large scale data. The future work not only considers local and global consistency in both involved spaces, but also is robust to the outliers and easy to handle the large scale issue.

REFERENCES

[1] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. 7th Int. Conf. KDD*, San Francisco, CA, USA, 2001, pp. 269–274.
[2] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proc. 9th Int. Conf. KDD*, Washington, DC, USA, 2003, pp. 89–98.
[3] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. 26th SIGIR*, Toronto, ON, Canada, 2003, pp. 267–273.
[4] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra, "Minimum sum-squared residue co-clustering of gene expression data," in *Proc. 4th ICDM*, 2004.
[5] B. Gao, T. Y. Liu, X. Zheng, Q. S. Cheng, and W. Y. Ma, "Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering," in *Proc. 11th Int. Conf. KDD*, Chicago, IL, USA, 2005, pp. 41–50.
[6] G. Qiu, "Image and feature co-clustering," in *Proc. 17th ICPR*, 2004, pp. 991–994.
[7] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, "A generalized maximum entropy approach to bregman co-clustering and matrix approximation," in *Proc. 10th Int. Conf. KDD*, New York, NY, USA, 2004, pp. 509–514.
[8] B. Gao, T.-Y. Liu, and W.-Y. Ma, "Star-structured high-order heterogeneous data co-clustering based on consistent information theory," in *Proc. 6th ICDM*, Washington, DC, USA, 2006, pp. 880–884.
[9] B. K. Bao, W. Min, K. Lu, and C. Xu, "Social event detection with robust high-order co-clustering," in *Proc. 3rd ACM ICMR*, Dallas, TX, USA, 2013, pp. 135–142.
[10] B. K. Bao, W. Min, J. Sang, and C. Xu, "Multimedia news digger on emerging topics from social streams," in *Proc. 20th ACM Int. Conf. Multimedia*, Nara, Japan, 2012, pp. 1357–1358.

[11] B. Long, Z. M. Zhang, and P. S. Yu, "Co-clustering by block value decomposition," in *Proc. 11th Int. Conf. KDD*, Chicago, IL, USA, 2005, pp. 635–640.
[12] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. 12th Int. Conf. KDD*, Philadelphia, PA, USA, 2006, pp. 126–135.
[13] J. S. Wu, J. H. Lai, and C. D. Wang, "A novel co-clustering method with intra-similarities," in *Proc. IEEE 11th ICDMW*, Vancouver, BC, Canada, 2011, pp. 300–306.
[14] L. Zhang *et al*., "Locally discriminative coclustering," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1025–1035, Jun. 2012.
[15] M. Wu and B. Scholkopf, "A local learning approach for clustering," in *Proc. Adv. NIPS*, 2007.
[16] W. Dai, G. R. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proc. 13th ACM SIGKDD*, San Jose, CA, USA, 2007, pp. 210–219.
[17] R. G. Pensa and J. F. Boulicaut, "Constrained co-clustering of gene expression data," in *Proc. ICDM*, 2008, pp. 25–36.
[18] D. Freitag, "Trained named entity recognition using distributional clusters," in *Proc. Conf. EMNLP-CoNLL*, 2004, pp. 262–269.
[19] R. Rohwer and D. Freitag, "Towards full automation of lexicon construction," in *Proc. Workshop CLS*, Stroudsburg, PA, USA, 2004, pp. 9–16.
[20] H. Li and N. Abe, "Word clustering and disambiguation based on co-occurrence data," in *Proc. 17th Int. Conf. COLING*, Stroudsburg, PA, USA, 1998, pp. 749–755.
[21] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, "A generalized maximum entropy approach to Bregman co-clustering and matrix approximation," *J. Mach. Learn. Res.*, vol. 8, pp. 1919–1986, Jan. 2007.
[22] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. NIPS*, 2004, pp. 321–328.
[23] J. Sun, Z. Shen, H. Li, and Y. Shen, "Clustering via local regression," in *Proc. ECML PKDD*, Antwerp, Belgium, 2008, pp. 456–471.
[24] F. Wang, C. Zhang, and T. Li, "Clustering with local and global regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 12, pp. 1665–1678, Dec. 2009.
[25] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2001.
[26] G. Strang, *Introduction to Linear Algebra*. Wellesly, MA, USA: Wellesley Cambridge, 2003.
[27] F. Bach and Z. Harchaoui, "DIFFRAC: A discriminative and flexible framework for clustering," in *Proc. NIPS*, 2007.
[28] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.
[29] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. 23rd SIGIR*, Athens, Greece, 2000, pp. 208–215.
[30] L. Lovász and M. D. Plummer, *Matching Theory*. Providence, RI, USA: AMS Chelsea Publishing, 2009.
[31] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Jan. 2003.
[32] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 50th Berkeley Symp. Math. Statist. Probab.*, Berkeley, CA, USA, 1967, pp. 281–297.
[33] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

Bing-Kun Bao received the Ph.D. degree in control theory and control application from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2009.

She is currently an Assistant Researcher with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, and a Researcher at China-Singapore Institute of Digital Media, Singapore. Her current research interests include cross-media cross-modal image search, social event detection, image classification and annotation, and sparse/low rank representation.

Dr. Bao received the Best Paper Award from ICIMCS'09, and served as a Technical Program Committee Member of several international conferences such as MMM2013, ICME2014, PCM2013. She also served as a Special Session Organizer in MMM2013 and PCM2013, and a Guest Editor for *Multimedia System Journal* and *Multimedia Tools and Applications*.

Weiqing Min received the B.E. degree from Shandong Normal University, Jinan, China, and the M.E. degree from Wuhan University, Wuhan, China, in 2008 and 2010, respectively.

He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was an Intern Student in the China-Singapore Institute of Digital Media, Singapore, in 2012. His current research interests include landmark-based multimedia search and mining, computer vision, and pattern recognition.

Teng Li received the B.E. degree in automation from the Department of Automation, University of Science and Technology of China, Hefei, China, and the M.E. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2010.

He was with the Institute of Automation, Chinese Academy of Sciences, in 2010. He is currently a Special-Term Professor (Associate Professor) with the College of Electrical Engineering and Automation, Anhui University, Anhui, China.

Changsheng Xu (M'97-SM'99-F'13) is a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and the Executive Director of China-Singapore Institute of Digital Media, Singapore.

His current research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He holds 30 granted/pending patents and has published over 200 refereed research papers in these areas.

Dr. Xu is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications and Applications*, and *ACM/Springer Multimedia Systems Journal*. He received the Best Associate Editor Award of *ACM Transactions on Multimedia Computing, Communications and Applications*, in 2012, and the Best Editorial Member Award of *ACM/Springer Multimedia Systems Journal*, in 2008. He has served as a Program Chair of ACM Multimedia in 2009. He has also served as an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair, and as a TPC Member for over 20 IEEE and *ACM Prestigious Multimedia Journals*, conferences, and workshops. He is an ACM Distinguished Scientist.