Simultaneous Discovery of Common and Discriminative Topics via Joint Nonnegative Matrix Factorization

Hannah Kim
Georgia Tech
hannahkim@gatech.edu

Jaegul Choo
Korea University
jchoo@korea.ac.kr

Jingu Kim
Netflix, Inc.
jingu.kim@gmail.com

Chandan K. Reddy
Wayne State University
reddy@cs.wayne.edu

Haesun Park
Georgia Tech
hpark@cc.gatech.edu

ABSTRACT

Understanding large-scale document collections in an efficient manner is an important problem. Usually, document data are associated with other information (e.g., an author's gender, age, and location) and their links to other entities (e.g., co-authorship and citation networks). For the analysis of such data, we often have to reveal common as well as discriminative characteristics of documents with respect to their associated information, e.g., male- vs. female-authored documents, old vs. new documents, etc. To address such needs, this paper presents a novel topic modeling method based on joint nonnegative matrix factorization, which simultaneously discovers common as well as discriminative topics given multiple document sets. Our approach is based on a block-coordinate descent framework and is capable of utilizing only the most representative, thus meaningful, keywords in each topic through a novel pseudo-deflation approach. We perform both quantitative and qualitative evaluations using synthetic as well as real-world document data sets such as research paper collections and nonprofit micro-finance data. We show our method has a great potential for providing in-depth analyses by clearly identifying common and discriminative topics among multiple document sets.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications- Data Mining; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Design, Performance

Keywords

Nonnegative matrix factorization; topic modeling; discriminative pattern mining

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783338.

1. INTRODUCTION

Topic modeling provides important insights from a large-scale document corpus [5, 16]. However, standard topic modeling does not fully serve the needs arising from many complex real-world applications, where we need to compare and contrast multiple document sets. For instance, such document sets can be generated as subsets of an entire data set by filtering based on their additional information, such as an author's gender, age, location, and relationships among these entities such as co-authorship and citation networks. Analyses on multiple document sets can provide interesting insights, especially when we can reveal the common or distinct characteristics among them. Another important application is time-evolving document analysis. Given recently published papers, it is often important to understand the currently emerging/diminishing research areas (distinct topics) and the research areas consistently studied over time (common topics).

For example, Fig. 1 shows the common topics and distinct topics between the research paper data sets from two different disciplines, namely, information retrieval and machine learning, produced by running the method proposed in this paper. A common topic between the two turns out to be language modeling based on a probabilistic framework such as hidden Markov models (Fig. 1(a)). On the other hand, information retrieval predominantly studies the topics about query expansion, database, and xml formats and the topics about the semantic web (Fig. 1(b)), while machine learning studies Bayesian approaches, neural networks, reinforcement learning, and multi-agent systems (Fig. 1(c)).

As another example, Fig. 2 shows the common topics and distinct topics among papers in the data mining area published in 2000-2005 and those published in 2006-2008, generated by our method. One can see that clustering and outlier/anomaly detection have been consistently studied over time (Fig. 2(a)). On the other hand, large-scale data mining and social network analysis have been recently emerging (Fig. 2(b)) while association rule mining and frequent pattern mining have received less attention during the later years (Fig. 2(c)).

We propose a novel topic modeling method that simultaneously discovers common topics and distinct topics out of multiple data sets, based on joint nonnegative matrix factorization (NMF). For simplicity, we focus on the case of two data sets. Nonnegative matrix factorization [23] has been widely used in document clustering and topic modeling [2, 3, 28, 30]. Our joint NMF-based topic modeling approach aims at simultaneously revealing common as well as distinct topics between two document sets, as shown in the pre-

(a) A common topic (b) Distinct topics for IR (c) Distinct topics for ML

Figure 1: The topic summaries of the research papers in information retrieval (IR) vs. those in machine learning (ML) disciplines.

(a) Common topics (b) Distinct topics for years 2006-2008 (c) Distinct topics for years 2000-2005

Figure 2: The topic summaries of the research papers in the data mining area published in 2000-2005 vs. those published in 2006-2008.

vious examples. We introduce two additional penalty terms into the objective function for joint NMF so that common topics can be as similar as possible between two data sets while the rest of the topics can be as different as possible. We also propose an approach where the dissimilarities among topics are defined only with the most representative keywords and show its advantages.

The main contributions of this paper are summarized as follows:

- We develop a joint NMF-based topic modeling method that simultaneously identifies common and distinct topics among multiple data sets.
- We perform a quantitative analysis using both synthetic and real-world document data sets, which shows the superiority of the proposed method.
- We show in-depth studies on various real-world document data sets including research paper collections as well as non-profit micro-finance/crowdfunding applications (available in Kiva.org).

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the problem formulation and our proposed NMF methods that incorporate commonality and distinctiveness penalty terms in the object function. Section 4 shows detailed quantitative results on both synthetic and real-world data sets. Section 5 presents the case studies of the proposed methods. Finally, Section 6 concludes our discussion.

2. RELATED WORK

There have been previous studies for simultaneously factorizing multiple data sets using NMF. Badea [4] considered NMF for two gene expression data sets with offset vectors. The primary objective was to identify common gene regulatory patterns, while constant expression levels in each data set are absorbed by the offset vectors. This approach is inadequate for more general use because an offset vector might be insufficient to describe uncommon aspects of data sets. Kim et al. [19] proposed group-sparsity regularization for NMF that can be used to discover common and different latent components from multiple data sets. Though their approach has the flexibility to discover the hidden structure of shared latent components, it does not explicitly incorporate the dissimilarities of unshared latent components, which is therefore not effective in contrasting data sets. Gupta et al. [15] imposed shared basis regularization on NMF for joint modeling of two related data sources. Their approach is less flexible than our method in that they set the shared space to be strictly identical. Simultaneous nonnegative factorization has also been used in different contexts such as knowledge transfer among multiple data sets [29], clustering that utilizes different views of a data set [25], and simultaneous clustering for multi-task learning [1].

Discriminative topic modeling, a variant of widely-used latent Dirichlet allocation [5], has been studied mainly to improve regression or classification performances [22, 31]. In [22, 31], class labels are utilized in order to form the latent topics so that the corresponding topical representations can be used to make accurate predictions. On the other hand, our primary goal is to discover the common and discriminative aspects of data sets, in order to gain better understanding of data. In [11], a topic modeling-based method has been proposed for generating updated summarization, with a goal to generate a compact summary of the new information, given a previously existing document set. Our work addresses a rather general scenario of comparing and contrasting multiple documents sets.

There exist other related research results in pattern mining [13], co-clustering [27], and network analysis [10]. Most of these results are based on heuristic approaches whereas ours is built on a theoretically sound framework. In addition, the existing studies are for the problems different from document analysis, and thus they are mostly limited to particular application domains such as bioinformatics. In this paper, we propose a more systematic NMF-based topic modeling approach that can extract discriminative topics from a wide variety of text data collections.

3. PROPOSED METHOD

In this section, we first introduce NMF in the topic modeling context. We then formulate our model and propose two methods, namely, the batch-processing approach (Section 3.3) and the pseudo-deflation approach (Section 3.4). The batch-processing approach produces common topics and discriminative topics by minimizing a single optimization criterion. The pseudo-deflation approach solves multiple optimization problems in a deflation-like manner and uses only the most representative keywords.

3.1 Preliminaries

Nonnegative matrix factorization (NMF) for document topic modeling. Given an input matrix $X \in \mathbb{R}_+^{m \times n}$, where $\mathbb{R}_+$ denotes the set of nonnegative real numbers, and an integer $k \ll \min(m, n)$, NMF solves a lower-rank approximation given by

$$X \approx WH^{\mathsf{T}}, \tag{1}$$

where $W \in \mathbb{R}_+^{m \times k}$ and $H \in \mathbb{R}_+^{n \times k}$ are factor matrices. This approximation can be achieved via various distance or divergence measures such as a Frobenius norm [17, 20], Kullback-Leibler divergence [23] and other divergences [12, 24]. Our methods are based on the widely-used Frobenius norm as follows:

$$\min_{W,H\geq 0} f(W,H) = \left\|X - WH^{\mathsf{T}}\right\|_F^2. \quad (2)$$

The constraints in the above equation indicate that all the entries of W and H are nonnegative. In topic modeling, $x_l \in \mathbb{R}_+^{m\times 1}$, the l-th column vector of X, corresponds to the bag-of-words representation of document l with respect to m keywords, possibly with some pre-processing, e.g., inverse-document frequency weighting and column-wise L_2-norm normalization. A scalar k corresponds to the number of topics. The l-th nonnegative column vector of W represents the l-th topic as a weighted combination of m keywords. A large value in a column vector of W indicates a close relationship of the topic to the corresponding keyword. The l-th column vector of H^{T} represents document l as a weighted combination of k topics, i.e., k column vectors of W.

3.2 Problem Formulation

Simultaneous common and discriminative topic modeling. Given a document set with n_1 documents and another document set with n_2 documents, our goal is to find $k(=k_c+k_d)$ topics from each document set, among which k_c topics are common between the two document sets and k_d topics are different between them.

Suppose we are given two nonnegative input matrices, $X_1 \in \mathbb{R}_+^{m\times n_1}$ and $X_2 \in \mathbb{R}_+^{m\times n_2}$, representing the two document sets and integers k_c and k_d. As shown in Fig. 3, we intend to obtain the NMF approximation of each input matrix as

$$X_1 \approx W_1 H_1^{\mathsf{T}} \text{ and } X_2 \approx W_2 H_2^{\mathsf{T}},$$

respectively, where $W_i = \left[\begin{array}{cc} W_{i,c} & W_{i,d}\end{array}\right] \in \mathbb{R}_+^{m\times k}$, $W_{i,c} \in \mathbb{R}_+^{m\times k_c}$, $W_{i,d} \in \mathbb{R}_+^{m\times k_d}$, and $H_i = \left[\begin{array}{cc} H_{i,c} & H_{i,d}\end{array}\right] \in \mathbb{R}_+^{n_i\times k}$ for $i = 1, 2$. We want to ensure the two topic sets for the common (or discriminative) topics represented as the column vectors of $W_{1,c}$ and $W_{2,c}$ (or $W_{1,d}$ and $W_{2,d}$) are as similar (or different) as possible.

We introduce two different penalty functions $f_c(\cdot,\cdot)$ and $f_d(\cdot,\cdot)$ for commonality and distinctiveness, respectively. A smaller value of $f_c(\cdot,\cdot)$ (or $f_d(\cdot,\cdot)$) indicates that a better commonality (or distinctiveness) is achieved. Using these terms, our problem is to optimize

$$\begin{aligned}\min_{W_1,H_1,W_2,H_2\geq 0} &\frac{1}{n_1}\left\|X_1 - W_1H_1^{\mathsf{T}}\right\|_F^2 + \frac{1}{n_2}\left\|X_2 - W_2H_2^{\mathsf{T}}\right\|_F^2 \\ &+\alpha f_c(W_{1,c},W_{2,c}) + \beta f_d(W_{1,d},W_{2,d}) \quad (3)\\ \text{subject to } &\|(W_1)_{\cdot l}\|_2 = 1, \|(W_2)_{\cdot l}\|_2 = 1 \text{ for } l = 1,\cdots,k,\end{aligned}$$

which indicates that, while performing lower-rank approximations on each of the two input matrices, we want to minimize both (1) the penalty for the commonality between the column set of $W_{1,c}$ and that of $W_{2,c}$ and (2) the penalty for the distinctiveness between the column set of $W_{1,d}$ and that of $W_{2,d}$. The coefficients $\frac{1}{n_1}$ and $\frac{1}{n_2}$ corresponding to the first and the second terms in Eq. (3) play a role of maintaining the balance between the different number of data items in X_1 and X_2. The parameters α and β control the weights of penalty functions for the approximation term. By solving this problem, we intend to reveal the common as well as the discriminative sets of topics between two data sets.

3.3 Batch-Processing Approach

To design an algorithm to solve Eq. (3), we first need to define $f_c(W_{1,c},W_{2,c})$ and $f_d(W_{1,d},W_{2,d})$. For algorithmic simplicity, we set them as

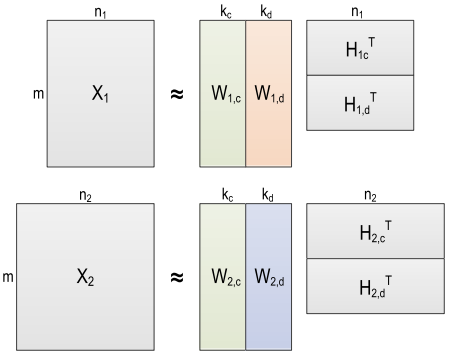


Figure 3: The illustration of our joint NMF-based topic modeling. Given the two term-document matrices, X_1 and X_2, the columns of $W_{1,c}$ and $W_{2,c}$ represent common topics while those of $W_{1,d}$ and $W_{2,d}$ represent the discriminative topics.

$$f_c(W_{1,c},W_{2,c}) = \left\|W_{1,c} - W_{2,c}\right\|_F^2 \text{ and} \quad (4)$$

$$f_d(W_{1,d},W_{2,d}) = \left\|W_{1,d}^{\mathsf{T}}W_{2,d}\right\|_{1,1}, \quad (5)$$

where $\|\cdot\|_{1,1}$ indicates the absolute sum of all the matrix entries. By plugging Eqs. (4)-(5) into Eq. (3), our overall objective function becomes

$$\begin{aligned}\min_{W_1,H_1,W_2,H_2\geq 0} &\frac{1}{n_1}\left\|X_1 - W_1H_1^{\mathsf{T}}\right\|_F^2 + \frac{1}{n_2}\left\|X_2 - W_2H_2^{\mathsf{T}}\right\|_F^2 \\ &+\alpha\left\|W_{1,c} - W_{2,c}\right\|_F^2 + \beta\left\|W_{1,d}^{\mathsf{T}}W_{2,d}\right\|_{1,1} \quad (6)\\ \text{subject to } &\|(W_1)_{\cdot l}\|_2 = 1, \|(W_2)_{\cdot l}\|_2 = 1 \text{ for } l = 1,\cdots,k.\end{aligned}$$

Using Eq. (4), we minimize the squared sum of element-wise differences between $W_{1,c}$ and $W_{2,c}$. In Eq. (5), the (i,j)-th component of $W_{1,d}^{\mathsf{T}}W_{2,d}$ corresponds to the inner product between $w_{1,d}^{(i)}$, the i-th topic vector of $W_{1,d}$, and $w_{2,d}^{(j)}$, the j-th topic vector of $W_{2,d}$. Thus, Eq. (5) represents the sum of the inner product values between all the possible column pairs between $W_{1,d}$ and $W_{2,d}$. By imposing the constraint $\|(W_i)_{\cdot l}\|_2 = 1$ and minimizing the sum of the absolute values, we encourage the sparsity in these inner products so that some of them become exactly zero. For any two nonnegative vectors $u, v \in \mathbb{R}_+^{m\times 1}$, their inner product $u^{\mathsf{T}}v = \sum_{p=1}^m u_p v_p$ is zero when for each p, either $u_p = 0$ or $v_p = 0$. Therefore, the penalty term based on Eq. (5) enforces each keyword to be related to only one topic, generating more discriminant topics representing differences between the two data sets.

Optimization. To solve Eq. (6), we propose an algorithm based on a block-coordinate descent framework that guarantees every limit point is a stationary point. We divide the set of elements in W and H, which are our variables to solve, into groups and iteratively solve each group while fixing the rest. First, we represent W_iH_i as the sum of rank-1 outer products [18], i.e.,

$$\begin{aligned}W_iH_i &= \sum_{l=1}^{k} w_i^{(l)}\left(h_i^{(l)}\right)^{\mathsf{T}} \quad (7)\\ &= \sum_{l=1}^{k_c} w_{i,c}^{(l)}\left(h_{i,c}^{(l)}\right)^{\mathsf{T}} + \sum_{l=1}^{k_d} w_{i,d}^{(l)}\left(h_{i,d}^{(l)}\right)^{\mathsf{T}} \text{ for } i = 1,2.\end{aligned}$$

where $w_i^{(l)}$, $h_i^{(l)}$, $w_{i,c}^{(l)}$, $h_{i,c}^{(l)}$, $w_{i,d}^{(l)}$, and $h_{i,d}^{(l)}$ represent the l-th column vectors of W_i, H_i, $W_{i,c}$, $H_{i,c}$, $W_{i,d}$, and $H_{i,d}$, respectively, and update these vectors one by one. By setting the derivatives of Eq. (3) to zero with respect to each of these vectors, we obtain the updating

rules as

$$w_{1,c}^{(l)} \leftarrow \left[\frac{\left(H_1^{\mathsf{T}} H_1\right)_{ll}}{\left(H_1^{\mathsf{T}} H_1\right)_{ll} + n_1 \alpha} w_{1,c}^{(l)} + \frac{X_1 h_{1,c}^{(l)} - W_1 \left(H_1^{\mathsf{T}} H_1\right)_{\cdot l} + n_1 \alpha w_{2,c}^{(l)}}{\left(H_1^{\mathsf{T}} H_1\right)_{ll} + n_1 \alpha} \right]_+, \quad (8)$$

$$w_{1,d}^{(l)} \leftarrow \left[w_{1,d}^{(l)} + \frac{X_1 h_{1,d}^{(l)} - W_1 \left(H_1^{\mathsf{T}} H_1\right)_{\cdot l} - n_1 \frac{\beta}{2} \sum_{p=1}^{k_d} w_{2,d}^{(p)}}{\left(H^{\mathsf{T}} H\right)_{ll}} \right]_+, \quad (9)$$

$$h_1^{(l)} \leftarrow \left[h_1^{(l)} + \frac{\left(X_1^{\mathsf{T}} W_1\right)_{\cdot l} - \left(H_1 W_1^{\mathsf{T}} W_1\right)_{\cdot l}}{\left(W_1^{\mathsf{T}} W_1\right)_{ll}} \right]_+, \quad (10)$$

where $[x]_+ = \max(x, 0)$ and $(\cdot)_{ll}$ represents the (l,l)-th component of a matrix in parentheses. After the update, $w_1^{(l)}$ is normalized to have a unit L_2-norm, and $h_1^{(l)}$ is multiplied correspondingly by $\left\| w_1^{(l)} \right\|_2$ for $l = 1, \cdots, k$. The updating rules for $w_{2,c}^{(l)}$, $w_{2,d}^{(l)}$, and $h_2^{(l)}$ can also be derived in a similar manner.

Computational Complexity. The proposed approach maintains the same complexity as the case of solving two separate standard NMF problems using the widely-used hierarchical alternating least squares (HALS) algorithm [9]. Both approaches follow the same block coordinate descent framework and require an equivalent computational cost for each iteration in this framework. In detail, updating $h_i^{(l)}$ is identical in both approaches, but the main difference lies in updating $w_i^{(l)}$ in which our approach has additional calculations as shown in the last terms of Eqs. (8)-(9). Since $H_i^T H_i$ and $X_i H_i$ can be pre-computed, the computational complexity of updating $w_i^{(l)}$ is $O(mk)$ in the standard NMF algorithm, where m is the number of keywords and k is the number of topics. In our approach, the computational complexity of updating $w_{i,c}^{(l)}$ is $O(mk)$ and that of updating $w_{i,d}^{(l)}$ is $O(m(k+k_d))$, which is still $O(mk)$ since $k = k_c + k_d$. Thus, the computational complexity of our approach for a single iteration of updating both $w_i^{(l)}$'s and $h_i^{(l)}$'s still remains the same as that of the standard NMF, i.e.,$O(mn_i k)$.

3.4 Pseudo-Deflation Approach

In this section, we first address several issues with the batch-processing algorithm from a practical standpoint and propose a novel method that considers only the most representative keywords in each topic. Similar to a rank-deflation procedure common in matrix factorization, this approach discovers discriminative topics one by one, hence the name "pseudo-deflation" approach.

The first point to discuss is that the penalty term for discriminative topics, as shown in Eq. (5), incorporates all the keywords (i.e., all m dimensions) when computing the inner product-based penalty value of two topic vectors. However, often in practice, only a small number of the most representative keywords are checked to understand the computed topics. Therefore, a better alternative would be to calculate the inner product in the penalty term using only the most representative keywords while ignoring the remaining insignificant keywords of each topic. Given a fixed number t, let $R_{1,d}^{(i)}$ and $R_{2,d}^{(j)}$ denote the sets of the t most representative keyword dimensions or indices from the two topic vectors, $w_{1,d}^{(i)}$ and $w_{2,d}^{(j)}$, respectively. Then, the (i,j)-th component of the penalty term for $f_d(W_{1,d}, W_{2,d})$ can be re-defined as

$$\left(f_d(W_{1,d}, W_{2,d})\right)_{ij} = \left(w_{1,d}^{(i)}\right)^{\mathsf{T}} I_m\left(R_{1,d}^{(i)} \cup R_{2,d}^{(j)}\right) w_{2,d}^{(j)} \quad (11)$$

where the diagonal matrix $I_m(S) \in \mathbb{R}_+^{m \times m}$ is defined as

$$\left(I_m(S)\right)_{pp} = \begin{cases} 1, & p \in S \\ 0, & p \notin S. \end{cases}$$

Note that $S \subset \{1, \cdots, m\}$ is a set of keyword dimensions/indices. We choose S as $R_{1,d}^{(i)} \cup R_{2,d}^{(j)}$ so that only the most representative keyword dimensions are used in the penalty function for distinctiveness.

Even though Eq. (11) provides more discriminative topics in terms of their most representative keywords, the main problem in using it in our joint NMF formulation is that the sets $R_{1,d}^{(i)}$ and $R_{2,d}^{(j)}$ can dynamically change as the intermediate results of topic vectors, $w_{1,d}^{(i)}$ and $w_{2,d}^{(j)}$, keep getting updated during algorithm iterations because a newly updated topic vector can have newly added/removed representative keywords. This causes our objective function, Eq. (3), itself to change over the iterations, and thus we can no longer guarantee that our algorithm monotonically improves the objective function value.

To overcome this issue, we now propose a pseudo-deflation-based approach that solves Eq. (3) incorporating Eq. (11). Our basic idea is to find discriminative topics in a greedy manner in order to keep the most representative keyword set of each topic fixed. In other words, we solve and fix one discriminative topic pair per stage. In the l-th stage, we find a discriminative topic pair $w_{1,d}^{(l)}$ and $w_{2,d}^{(l)}$ that are distinct from the discriminative topics obtained from the previous stages, $\{w_{2,d}^{(1)}, \cdots, w_{2,d}^{(l-1)}\}$ and $\{w_{1,d}^{(1)}, \cdots, w_{1,d}^{(l-1)}\}$ respectively, and are different from each other. As a result, the entire solution is discovered after k_d stages.

The proposed approach is outlined as follows: First, given the two input matrices X_1 and X_2 and integers k_c and k_d, we set

$$k_c^s = k_c + k_d = k \text{ and } k_d^s = 0$$

where k_c^s (or k_d^s) is the temporarily assigned number of common (or discriminative) topics at each stage, and solve Eq. (12). We first attempt to find k common topics of X_1 and X_2 in the first stage. In the next stage, we decrease k_c^s and increase k_d^s by 1 (to find $k-1$ common topics and 1 discriminative topic) and solve a new objective function

$$\begin{aligned} \min_{W_{1,c}, w_{1,d}^{(k_d^s)}, H_1, W_{2,c}, w_{2,d}^{(k_d^s)}, H_2 \geq 0} & \frac{1}{n_1} \left\| X_1 - W_1 H_1^{\mathsf{T}} \right\|_F^2 + \frac{1}{n_2} \left\| X_2 - W_2 H_2^{\mathsf{T}} \right\|_F^2 \\ & + \frac{\alpha}{k_c^s} \sum_{l=1}^{k_c^s} \left\| w_{1,c}^{(l)} - w_{2,c}^{(l)} \right\|_2^2 + \frac{\beta}{k_d^s - 1} \sum_{l=1}^{k_d^s - 1} \left(w_{1,d}^{(l)}\right)^{\mathsf{T}} I_m\left(R_{1,d}^{(l)}\right) w_{2,d}^{(k_d^s)} \\ & + \frac{\beta}{k_d^s - 1} \sum_{l=1}^{k_d^s - 1} \left(w_{2,d}^{(l)}\right)^{\mathsf{T}} I_m\left(R_{2,d}^{(l)}\right) w_{1,d}^{(k_d^s)} + \gamma \left(w_{1,d}^{(k_d^s)}\right)^{\mathsf{T}} \left(w_{2,d}^{(k_d^s)}\right) \quad (12) \\ & \text{subject to } \|(W_1)_{\cdot l}\| = 1, \|(W_2)_{\cdot l}\| = 1 \text{ for } l = 1, \cdots, k. \end{aligned}$$

We progressively solve this equation after decreasing k_c^s and increasing k_d^s values by one until k_d^s becomes k_d.

When solving Eq. (12), we fix $w_{i,d}^{(l)}$'s for $i = 1, 2$ and $l = 1, \cdots, k_d^s - 1$ as those obtained from previous stages, and solve only the rest of the topics in W_i. In this manner, each pair of discriminative topics $w_{1,d}^{(l)}$ and $w_{2,d}^{(l)}$ is determined one by one and is fixed throughout the subsequent stages that use different k_c^s and k_d^s values. Notice that a typical successive rank-1 deflation method, which is common in singular value decomposition, e.g., the power iteration [14], does not guarantee an optimal solution for NMF [18]. For example, the basis vector obtained by a rank-1 NMF is not necessarily part of

those obtained by a rank-2 NMF, and they can be quite different. To effectively handle this problem, our approach maintains the same number of topics throughout the progression of stages while a subset of the basis vectors are fixed. In this respect, we call our method a pseudo-deflation approach.

The main advantage of such a pseudo-deflation approach is the ability to maintain the fixed set of the representative keyword indices. By fixing $w_{i,d}^{(l)}$'s from the previous stages, we can now maintain the constant set $R_{i,d}^{(l)}$'s for them in the penalty terms for distinctiveness, as shown in the fourth and the fifth terms in Eq. (12), which makes the objective function remain the same over iterations within a single stage. Finally, the last term in Eq. (12) plays a role of enhancing the distinction between the topic pair $w_{1,d}^{(k_d^s)}$ and $w_{2,d}^{(k_d^s)}$. Nonetheless, $w_{i,d}^{(k_d^s)}$ can still have a varying set $R_{i,d}^{(k_d^s)}$ during iterations, and thus we just use the inner product over the entire set of dimensions.

Parameter adaptation. The proposed pseudo-deflation method contains various elements contributing to the penalty terms for commonality and distinctiveness while k_c^s and k_d^s change. Thus, unlike the parameters α and β in Eq. (6), we adaptively change the regularization parameters so that the total penalty values are comparable among various k_c^s and k_d^s values. Therefore, the penalty terms of Eq. (12) contain denominators as the number of total contributing columns for each penalty term.

Optimization. Eq. (12) can be optimized in a similar manner shown in Eqs. (8)-(10) based on the block-coordinate descent framework. The updating rules can be described as

$$w_{1,c}^{(l)} \leftarrow \left[\frac{\left(H_1^\mathsf{T} H_1\right)_{ll}}{\left(H_1^\mathsf{T} H_1\right)_{ll} + n_1 \alpha} w_{1,c}^{\{l\}} + \frac{X_1 h_{1,c}^{\{l\}} - W_1 \left(H_1^\mathsf{T} H_1\right)_{\cdot l} + n_1 \frac{\alpha}{k_c^s} w_{2,c}^{\{l\}}}{\left(H_1^\mathsf{T} H_1\right)_{ll} + n_1 \alpha} \right]_+, \tag{13}$$

$$w_{1,d}^{(k_d^s)} \leftarrow \left[w_{1,d}^{(k_d^s)} + \frac{X_1 h_{1,d}^{(k_d^s)} - W_1 \left(H_1^\mathsf{T} H_1\right)_{\cdot k_d^s}}{\left(H^\mathsf{T} H\right)_{k_d^s k_d^s}} - \frac{n_1 \frac{\beta}{2(k_d^s - 1)} \sum_{p=1}^{k_d^s - 1} I\left(R_{2,d}^{(p)}\right) w_{2,d}^{(p)} + n_1 \frac{\gamma}{2} w_{2,d}^{(k_d^s)}}{\left(H^\mathsf{T} H\right)_{k_d^s k_d^s}} \right]_+, \tag{14}$$

and the same updating rule applies for $h_1^{(l)}$ as in Eq. (10). After the update, $w_1^{(l)}$ is normalized to have a unit L_2-norm, and $h_1^{(l)}$ is multiplied correspondingly by $\left\| w_1^{(l)} \right\|_2$ for $l = 1, \cdots, k$. The updating rules for $w_{2,c}^{(l)}$, $w_{2,d}^{(l)}$, and $h_2^{(l)}$ can also be derived in a similar manner. Finally, our algorithm is summarized in Algorithm 1.

Initialization. A single stage inside the for-loop in Algorithm 1 can be considered as introducing an additional pair of discriminative topics between two data sets while removing a common topic pair, as k_c^s and k_d^s get updated. In this process, it is important to provide a capability to maintain a consistent result set and smooth transition. To this end, we use the following initialization strategy for Eq. (12). Given a result set for particular values of k_c^s and k_d^s, we choose a common topic pair that has the lowest approximation capability for input matrices and set them as the initial discriminative topic pair for the next stage, i.e.,

$$\arg\min_{w_{1,c}^{(l)}, w_{2,c}^{(l)}} \sum_{i=1}^{2} \left\| w_{i,c}^{(l)} \left(h_{i,c}^{(l)}\right)^\mathsf{T} - \max\left(w_{i,c}^{(l)} \left(h_{i,c}^{(l)}\right)^\mathsf{T} - X_i, 0_{m \times n} \right) \right\|_F^2, \tag{15}$$

where the max operation applies in an element-wise manner.

Algorithm 1: The Pseudo-deflation-based joint NMF

Input: Two input matrices X_1 and X_2, integers k_c and k_d, and parameters α, β, and γ

Output: $W_i = \left[\begin{array}{cc} W_{i,c} & W_{i,d} \end{array} \right] \in \mathbb{R}_+^{m \times k}$ and $H_i = \left[\begin{array}{cc} H_{i,c} & H_{i,d} \end{array} \right] \in \mathbb{R}_+^{n_i \times k}$ for $i = 1, 2$

Initialize W_i and H_i for $i = 1, 2$;

for $k_d^s \leftarrow 0$ **to** k_d **do**

  $k_c^s \leftarrow k_c + k_d - k_d^s$;

  `/* For` k_c^s `and` k_d^s`, solve Eq. (12) */`

  repeat

    Update W_i's using Eqs. (13)-(14);

    Update H_i's using Eq. (10);

    Normalize columns of W_i's to have unit norms and scale H_i's accordingly;

  until *a stopping criteria is satisfied*;

  Choose $w_{1,c}^l$ and $w_{2,c}^l$ satisfying Eq. (15);

  `/* Remove` $w_{i,c}^l$ `from` $W_{i,c}$ `*/`

  $W_{i,c} \leftarrow W_{i,c} \backslash w_{i,c}^l$ for $i = 1, 2$;

  `/* Append` $w_{i,c}^l$ `to` $W_{i,d}$ `on the right side */`

  $W_{i,d} \leftarrow \left[\begin{array}{cc} W_{i,d} & w_{i,c}^l \end{array} \right]$ for $i = 1, 2$;

end

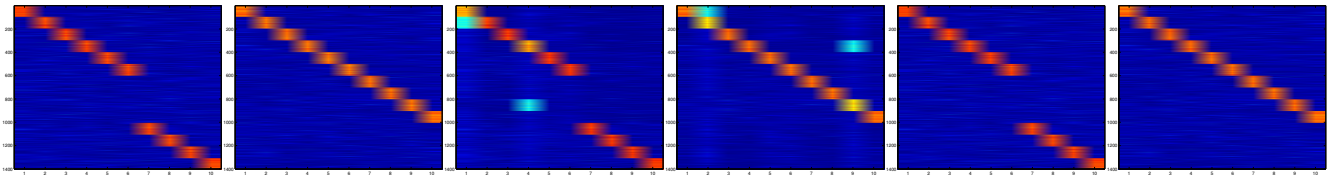
Computational Complexity. Similar to the batch processing approach, the pseudo-deflation approach involves additional computations (the last term of Eq. (13) and the last two terms of Eq. (14)) in the updating step of $w_i^{(l)}$ compared to the standard NMF. Since $H_i^T H_i$ and $X_i H_i$ can be pre-computed, the computational complexity of updating $w_{i,c}^{(l)}$ is $O(mk)$ and that of updating $w_{i,d}^{(k_d^s)}$ is $O(mk + tk_d^s)$, where t is the number of top keyword, which then becomes equivalent to $O(mk)$ since the number of top keywords of our interest, t, is relatively small. Therefore, the overall complexity of the pseudo-deflation approach for an iteration of updating both $w_i^{(l)}$'s and $h_i^{(l)}$'s, is still $O(mn_ik)$, which is the same as that of the standard NMF. Note that the complexity of the pseudo-deflation approach is approximately k_d times that of the batch processing approach since it solves Eq. (12) in k_d stages. However, the problem size decreases as the stage progresses since the pseudo-deflation approach do not solve for $w_{i,d}^{(l)}$'s that are already obtained from the previous stages.

4. QUANTITATIVE EVALUATION

In this section, we evaluate our proposed methods using synthetic as well as various real-world data sets. First, we present quantitative results on synthetic data to show the superiority of the pseudo-deflation method against the batch-processing method. We then provide the results of our methods using real-world data sets and compare them with other alternative solutions.

4.1 Basis Reconstruction on Synthetic Data

We conduct analysis on a synthetic data set and compare the batch-processing approach and the pseudo-deflation approach.



(a) Ground-truth matrices (b) The batch-processing method (c) The pseudo-deflation method

Figure 4: Ground-truth matrices for W_1(left) and W_2(right) and their computed results by the two proposed approaches.

4.1.1 Data Generation

We apply our proposed methods to a synthetic data set for which the ground-truth factor matrices are known. We generate the two input matrices, $X_i \in \mathbb{R}_+^{1600\times 300}$ for $i = 1, 2$, which can be considered as term-document matrices based on their factor matrices $W_{i,c} = \left[\begin{array}{ccc} w_{i,c}^{(1)} & \cdots & w_{i,c}^{(6)} \end{array}\right] \in \mathbb{R}_+^{1600\times 6}$ and $W_{i,d} = \left[\begin{array}{ccc} w_{i,d}^{(1)} & \cdots & w_{i,d}^{(4)} \end{array}\right] \in \mathbb{R}_+^{1600\times 4}$ as

$$\left(w_{i,c}^{(l)}\right)_p = \begin{cases} 1, & 100\,(l-1) < p \le 100l \\ 0, & \text{otherwise} \end{cases}, \text{ and}$$

$$\left(w_{i,d}^{(l)}\right)_p = \begin{cases} 1, & idx(i,l) < p \le idx(i,l) + 100 \\ 0, & \text{otherwise} \end{cases},$$

where $idx(i,l) = 600 + 400\,(i-1) + 100\,(l-1)$. In other words, each of the six common topic pairs, $w_{1,c}^{(l)}$ and $w_{2,c}^{(l)}$, contains nonzero elements in 100 common dimensions while the four discriminative topic pairs (eight in total) have 100 nonzero entries in a completely disjoint dimension set. In addition, each row of $H_i \in \mathbb{R}_+^{300\times 10}$ is set to be a unit vector that has only one nonzero entry at a randomly selected dimension. Afterwards, we add a random Gaussian noise to each entry of W_i and H_i and form X_i as the product of them, $W_i H_i^{\mathsf{T}}$, with an additional random Gaussian noise added to each element of them.

4.1.2 Results

Fig. 4(a) shows the resulting ground-truth matrices for W_1 (left) and W_2 (right). Figs. 4(b) and 4(c) show the examples of the resulting W_1 (left) and W_2 (right), which are computed by the batch-processing and the pseudo-deflation methods, respectively. As can be seen in these figures, the latter successfully reconstructs the ground-truth matrices while the batch-processing method does not. To test our claim, we run each algorithm 20 times with random initializations while providing identical initializations to both algorithms at each run. Fig. 5 shows the reconstruction error of X_i over 20 runs of each algorithm with different k_d's. As expected, both methods show minimum reconstruction error when k_d is set to 6, which is the correct number of discriminative topic pairs. The pseudo-deflation method consistently outperforms the batch-processing approach in terms of a reconstruction error with a much smaller variance. In the results, this indicates that *the pseudo-deflation method is less susceptible to noise in the data and it gives more consistent results that are closer to the true solution among multiple runs*.

4.2 Algorithimic Evaluation

4.2.1 Experimental Setup

To analyze the behavior of our proposed methods, we use the following real-world document data sets with different partitionings: VAST-InfoVis papers published in the two closely related IEEE conferences in the field of visualization, namely, Visual Analytics Science and Technology (VAST) and Information Visualization

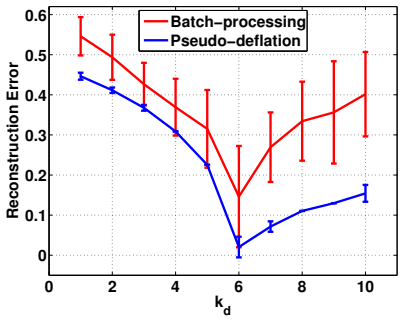


Figure 5: The reconstruction error ($\sum_{i=1}^{2} \frac{1}{n_i} \left\|X_i - W_i H_i^{\mathsf{T}}\right\|_F^2$) vs. k_d for the synthetic data. The results are averaged over 20 runs, and the error bar represents their variance. k (=$k_c + k_d$) is set to 10.

(InfoVis) (2 groups, 515 documents, 5,935 keywords),[1] and Four Area paper data published in machine learning (ML), databases (DB), data mining (DM), and information retrieval (IR) fields (4 groups, 15,110 documents, 6,487 keywords).[2]

For each pair of data sets (one pair for VAST-InfoVis data set and six pairs for Four Area data set), we evaluated the quality of the topic modeling results in terms of three different measures: the reconstruction error, the distinctiveness score, and the commonality score. The reconstruction error is defined as the sum of the first two terms in Eq. (3) (See the caption of Fig. 5). The commonality score is defined as Eq. (4) divided by the number of common topics, k_c, indicating how close common topics $w_1^{(l)}$'s are to their corresponding common topics $w_2^{(l)}$'s in the other document set. Finally, we use the distinctiveness score as an averaged symmetrized Kullback-Leibler divergence between all the discriminative topic pairs, i.e.,

$$\frac{1}{2k_d^2} \sum_{i=1}^{k_d} \sum_{j=1}^{k_d} \left[\left(w_{1,d}^{(i)}\right)^{\mathsf{T}} \log(w_{1,d}^{(i)}) + \left(w_{2,d}^{(j)}\right)^{\mathsf{T}} \log(w_{2,d}^{(j)}) - \left(w_{1,d}^{(i)}\right)^{\mathsf{T}} \log(w_{2,d}^{(j)}) - \left(w_{2,d}^{(j)}\right)^{\mathsf{T}} \log(w_{1,d}^{(i)}) \right], \quad (16)$$

which indicates how distinct the obtained discriminative topics are. For the first measure, a lower value indicates a better quality while a higher value indicates a better quality for the second and the third measures.

We compared three different methods: (1) the standard NMF, (2) the batch-processing method, and (3) the pseudo-deflation method. For the first one, after obtaining the two topic sets by applying NMF separately to each of the two sets, we choose k_c topic pairs that have the highest commonality scores and treat them as the common topic pairs and the rest as the discriminative ones. For parameters to be specified to run the batch-processing method (Eq. (6)) and the

[1] `http://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html`

[2] `http://dais.cs.uiuc.edu/manish/ECOutlier/`

Table 1: The evaluation results based on three different measures on real-world data sets. The reported results are averaged values over 20 runs. The best performance values are shown in bold.

Data sets	Reconstruction error			Commonality score			Distinctiveness score		
	Standard NMF	Batch processing	Pseudo-deflation	Standard NMF	Batch processing	Pseudo-deflation	Standard NMF	Batch processing	Pseudo-deflation
VAST-InfoVis	**1.7116**	1.7804	1.7409	.3611	**.0011**	.0041	206.1248	188.9593	**239.1429**
Four Area (ML-DB)	**.0705**	.0712	.0710	.4409	.0011	**.0003**	108.0325	105.8713	**121.1697**
Four Area (ML-DM)	**.0737**	.0746	.0758	.3206	.0007	**.0005**	111.9828	117.7371	**119.3134**
Four Area (ML-IR)	**.0717**	.0726	.0725	.3162	.0012	**.0005**	105.8652	104.1647	**116.0636**
Four Area (DB-DM)	**.0778**	.0791	.0787	.4412	.0013	**.0004**	95.6500	109.4650	**110.2718**
Four Area (DB-IR)	**.0758**	.0771	.0764	.2635	.0012	**.0004**	96.1121	99.8529	**103.6566**
Four Area (DM-IR)	**.0790**	.0802	.0800	.2905	.0011	**.0004**	87.5875	97.4784	**103.6090**

pseudo-deflation method (Eq. (12)), we adaptively set them to be sufficiently large so that no common keywords occur among the ten most representative keywords between discriminative topics from different data sets. At the same time, we make sure that the ten most representative keywords between common topic pairs become identical.

4.2.2 Results

Table 1 shows the quantitative comparisons among different methods with respect to various measures. It is not surprising to see that the standard NMF achieves the lowest reconstruction errors for all the cases since its objective is entirely to minimize the reconstruction error. However, its commonality as well as discriminative scores are shown to be significantly lower compared to the two other methods, which implies the limitation of the standard NMF for comparison/contrasting purposes.

The reconstruction errors of the two other methods are comparable to the standard NMF results. For all the cases except for the ML-DB case in the Four Area data set, the pseudo-deflation method shows better reconstruction errors than the batch-processing method, but at the same time, the former generally performs better than the latter in terms of both the commonality and the discriminative scores, as seen in all the Four Area data cases. These observations are consistent with the previous results using the synthetic data set (Section 4.1), which highlights the advantage of the pseudo-deflation method over the batch-processing method.

4.3 Clustering Performance

We now apply our method for clustering of real-world data sets. We assume that multiple data sets share common clusters while each of them has its own exclusive clusters. Our hypothesis here is that by jointly performing clustering on multiple data sets allowing both common and discriminative topics, our method will have advantages over other methods that perform independent clustering on each data set and other joint NMF-based methods [15, 25].

4.3.1 Experimental Setup

To evaluate our method in clustering applications, we used various real-world document data sets: 20 Newsgroup data (20 clusters, 18,828 documents, 43,009 keywords),[3] Reuters data (65 clusters, 8,293 documents, 18,933 keywords),[4] and Four Area data set described in Section 4.2. All these data sets are encoded as term-document matrices using term frequency values, and for each data set, we formed two document subsets as shown in Table 2. We note that even though the two subsets have common clusters, we randomly split the data items in such clusters to the two subsets so that no data items overlap between them.

We compared the three following methods to our methods (batch-processing and pseudo-deflation approaches): (1) the standard NMF, which is applied separately to each subset, (2) Multi-View NMF (MV-NMF) [25], and (3) regularized shared subspace NMF (RS-NMF) [15]. For MV-NMF, the problem setting assumes the input data sets are two different representations of a single data set, whereas our method assumes that the two different data sets are represented in the same feature space. To resolve this discrepancy, we used the transposed version of MV-NMF so that it can be applied in our setting.

The parameters used in the following experiments are as follows. For the batch processing method, we set parameter α (in Eq. (6)) as 100 and parameter β (in Eq. (6)) as 10. For the pseudo-deflation method, we set parameter α (in Eq. (12)) as 100 and parameters β and γ (in Eq. (12)) as 10, but we found that our method is not sensitive to these parameter values. For MV-NMF, we used the default setting in their implementation available on the author's webpage.[5] For RS-NMF, we used a common weighting parameter a as 100, as suggested in [15]. We used the identical initialization for all the compared methods.

4.3.2 Results

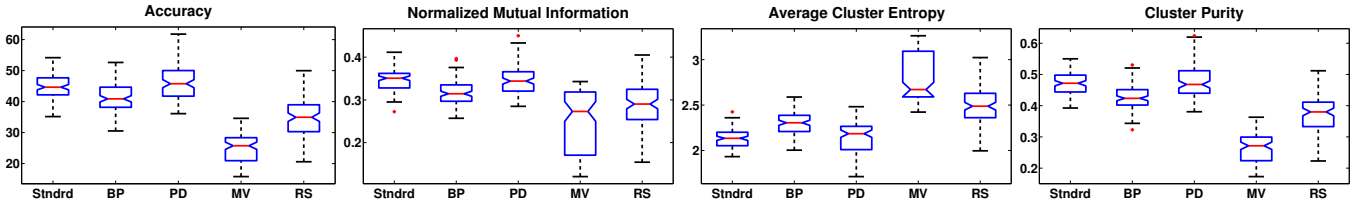
Our experiments tested how well the computed clustering outputs match the ground-truth cluster labels. We first computed the cluster index of each data item as the most strongly associated topic index based on its corresponding column vector of H_i. Next, we re-mapped the obtained cluster indices to the ground-truth labels using the Hungarian algorithm [21]. Then, we applied four widely-adopted cluster quality measures to the computed cluster indices: accuracy, normalized mutual information, averaged cluster entropy, and cluster purity [26].

Fig. 6 shows these results from 100 runs of each case. In all the results, our methods, batch processing and pseudo-deflation approaches, outperform existing methods such as MV-NMF and RS-NMF in all the four measures. The reason for inferior performance of MV-NMF is because it aims to find only common topics and do not consider discriminative topics. On the other hand, RS-NMF can take into account both common as well as discriminative topics but its main drawback is the lack of flexibility since it imposes the common topics strictly to be the same across multiple data sets. Between the batch processing and the pseudo-deflation method, the latter generally shows better performances than the former except for the accuracy measure from the Four Area data set (Fig. 6(c)). This shows the superiority of our carefully designed pseudo-deflation method in practical applications.

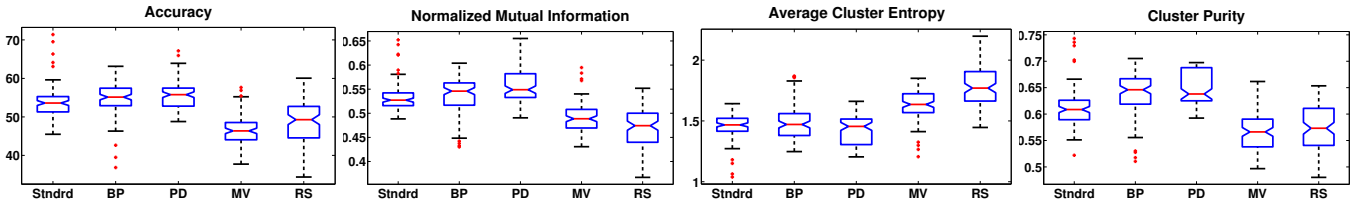
[3] http://qwone.com/~jason/20Newsgroups/

[4] http://www.cc.gatech.edu/~hpark/othersoftware_data.php

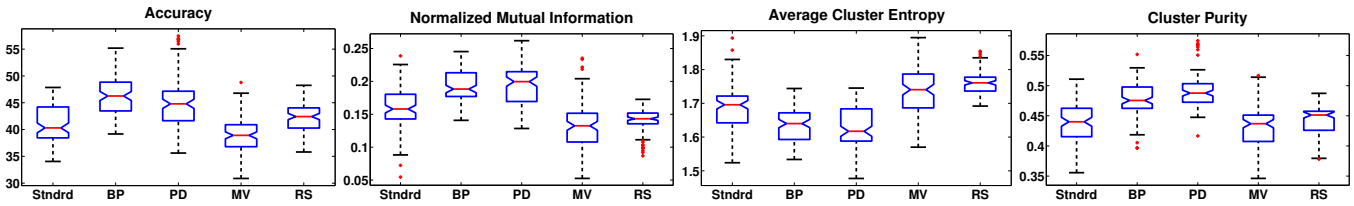
[5] http://jialu.cs.illinois.edu/code/Code_multiNMF.zip



(a) 20 Newsgroup data set



(b) Reuters data set



(c) Four Area data set

Figure 6: The clustering performance of the standard NMF (Stndrd), batch processing method (BP), pseudo-deflation method (PD), MV-NMF (MV), and RS-NMF (RS) measured in terms of accuracy, normalized mutual information, cluster entropy, and cluster purity metrics. For each case, 100 repetitive runs with different random initializations were used. Higher values indicate better performance except for clustering entropy.

Table 2: The clusters contained in document subsets.

	Common clusters	Exclusive clusters in subset 1	Exclusive clusters in subset 2
20 News-group	'comp.windows.x', 'sci.med', 'sci.space', 'soc.religion.christian'	'alt.atheism', 'rec.sport.baseball', 'sci.electronics', 'talk.politics.guns'	'comp.sys.mac.hardware', 'rec.sport.hockey', 'sci.crypt', 'talk.politics.mideast'
Reuters	'sugar', 'gnp', 'cpi'	'crude', 'interest', 'coffee', 'gold', 'reserves'	'trade', 'money-fx', 'ship', 'money-supply', 'cocoa'
Four Area	data mining, information retrieval	machine learning	database

5. TOPIC DISCOVERY EXAMPLES

Previously, we evaluated our method in terms of computing the true low-rank factors as well as jointly clustering multiple data sets. In this section, we discuss the meaningful topics that our method can discover in various applications, which can broaden our insights about the data. In Figs. 7-10, the results are visualized using Wordle[6] based on the weight values of the basis vectors.

5.1 VAST vs. InfoVis Conference Papers

The first case study is performed on VAST-InfoVis data set described in Section 4.2. As shown in Fig. 7, the two venues share the common topics of interactive visualization techniques and user interface systems. On the other hand, the topics studied exclusively in VAST are shown to be decision making processes as well as high-dimensional data visualization using clustering and dimension reduction, e.g., the paper "Similarity clustering of dimensions for an enhanced visualization of multidimensional data" by Ankerst et al. The exclusive topics in InfoVis include graph drawing/layout algorithms and color blending/weaving techniques, e.g., the paper "Weaving Versus Blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color" by Hagh-Shenas et al.

[6]http://www.wordle.net

5.2 Loan Description Data in Micro-finance

Next, we apply our methods to the text data available in a novel domain of micro-finance at Kiva.org.[7] Kiva.org is a non-profit website where people in developing countries, who lack access to financial services for their economic sustainability, can post a loan request and in response, other people can easily lend a small amount of money to them in a crowd-funding framework. Lending activities are entirely driven by altruism since the lenders do not gain any financial profit or interest, and thus it is crucial to understand people's lending behaviors in order to increase lending activities and help people sustain their lives. Even with such a social impact of this domain, only a little research has been conducted so far [6, 8].

Kiva.org contains rich textual data and other information associated with them. For example, a loan request is available in a free text form, and it describes the borrower and the purpose of a loan, Additionally, there exists various information about other associated entities such as lenders, lending teams (a group of lenders with a common interest), and field partners (those helping borrowers with the loan terms and conditions) in terms of their ages, genders, occupations, and geo-locations, etc. By analyzing a set of textual descriptions of the loans that particular groups of lenders (e.g., with the same location, occupation, etc.) or lending teams

[7]The processed data is available at http://fodava.gatech.edu/kiva-data-set-preprocessed

(a) Common topics (b) Distinct topics for VAST (c) Distinct topics for InfoVis

Figure 7: The topic summaries of the research papers published in VAST vs. InfoVis conferences.

(a) Common topics (b) Distinct topics for 'Etsy.com Handmade' (c) A distinct topic for 'Guys holding fish'

Figure 8: The topic summaries of the loans funded by the lending teams 'Guys holding fish' vs. 'Etsy.com Handmade'.

have funded, our method can help to characterize their lending behaviors, which will then be utilized in increasing lending activities. In the following, we describe several examples of such in-depth analyses.

Lending Teams 'Etsy.com Handmade' vs. 'Guys holding fish'. First, we choose the two interesting lending teams, 'Etsy.com Handmade' and 'Guys holding fish', and analyze the common as well as the distinct topics in the textual descriptions of the loans that each team funded. Fig. 8 shows their common topics as those loans related to buying products such as food and clothes in order to resell them in his/her stores as well as farming-related needs including buying seeds and fertilizers. On the other hand, the former team 'Etsy.com Handmade', which consists of the users of an online marketplace for handcrafted items, shows distinct characteristics of funding the loans related to fabric, threads, beads, and sewing machines as well as those related to clothes and shoes, e.g., the loans to buy more fabrics, threads, and laces for tailoring business. The team 'Guys holding fish' tends to fund loans related to fishing, e.g., buying/repairing fishing equipment such as boats, fishing nets, and other gears. These interesting behaviors can be expressed as *homophily*, as observed by the fact that *people tend to fund loans similar to what they like*.

Lending Teams 'Thailand' vs. 'Greece'. Next, we choose the two lending teams based on their geo-location, 'Thailand' and 'Greece', and analyze their topics in the loans they fund. As shown in Fig. 9, the common loans that both teams fund are related to buying groceries and supplying stock for borrowers' stores as well as expanding borrowers' business via investment. On the other hand, the former team particularly funds the loans related to buying ingredients such as vegetable, fruit, meat, oil, sugar, rice, and flour in order to resell or use them in cooking business. However, the latter focuses on the loans related to purchasing materials such as cement, stone, sand, or paint for construction business as well as buying furniture/appliances for home improvement or for shops. Interestingly, according to the World Bank, about 40 percent of Thailand laborers work in agriculture while only 13 percent of Greece employment is in agriculture. We also found that construction and manufacturing industrial products such as cement and concrete are the two main industries in Greece. This finding shows another example of homophily in lending behaviors.

Lender Occupations. Finally, we generated the loan subsets that were funded by lenders characterized by their occupations. To this end, we first formed groups of lenders whose occupation description fields contain a particular keyword. Next, we generated the subset of loans associated with this lender group. Then, we performed our topic analysis on this loan subset against a set of randomly selected loans. Fig. 10 shows several examples of distinct topics that such a lender group is associated with. For instance, those lenders with 'art'-related occupations like to fund the loans related to buying and selling clothes, shoes, and cosmetics as well as purchasing material related to sewing and house construction, in contrast to random lenders. Another lender group associated with the occupation 'driver' likes to fund the loans related to buying, repairing, or maintaining vehicles such as taxis, motorcycles, and trucks. Finally, the lender group associated with the occupation 'teacher' is clearly shown to fund school-related loans such as paying fees and tuitions for children's schools, universities, and trainings.

6. CONCLUSION

In this paper, we proposed a joint topic modeling approach based on nonnegative matrix factorization that supports the needs to compare and contrast multiple data sets. To solve our novel NMF-based formulation, we utilized a block-coordinate descent framework based on a rank-one outer product form and proposed the novel pseudo-deflation method, which takes into account only the most representative keywords. For our evaluation, we provided detailed quantitative analysis using both synthetic and real-world document data, which shows the superiority of our proposed methods. We also provided in-depth analyses for comparing and contrasting various document data in the context of research paper data as well as non-profit micro-finance activity data. Through these quantitative and qualitative analyses, our experiments show that the proposed approach clearly identifies common and distinct topics that provide a deep understanding when handling multiple document data sets.

As our future work, we plan to improve the efficiency of the proposed methods so that they can support on-the-fly real-time computations given dynamically filtered document subsets. In addition, we plan to build a visual analytics system where the computed common and discriminative topics are interactively visualized along with their associated documents [7].

Acknowledgments

This work was supported in part by DARPA XDATA grant FA8750-12-2-0309 and NSF grants CCF-1348152, IIS-1242304, and IIS-1231742. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of funding agencies.

7. REFERENCES

[1] S. Al-Stouhi and C. K. Reddy. Multi-task clustering using constrained symmetric non-negative matrix factorization. In *Proc.*

(a) Common topics (b) Distinct topics for 'Thailand' (c) Distinct topics for 'Greece'

Figure 9: The topic summaries of the loans funded by the lending teams 'Thailand' vs. 'Greece'.

(a) The keyword 'art' (b) The keyword 'driver' (c) The keyword 'teacher'

Figure 10: Distinct topics computed from the loans funded by those whose occupations contain a particular keyword. Another loan group is set to a set of randomly selected loans.

SIAM International Conference on Data Mining (SDM), pages 785–793, 2014.

[2] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. *Journal of Machine Learning Research (JMLR)*, 28(2):280–288, 2013.

[3] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization – provably. In *Proc. the 44th Symposium on Theory of Computing (STOC)*, pages 145–162, 2012.

[4] L. Badea. Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In *Proc. the Pacific Symposium on Biocomputing*, pages 267–278, 2008.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

[6] J. Choo, C. Lee, D. Lee, H. Zha, and H. Park. Understanding and promoting micro-finance activities in kiva.org. In *Proc. the 7th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 583–592, 2014.

[7] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):1992–2001, 2013.

[8] J. Choo, D. Lee, B. Dilkina, H. Zha, and H. Park. A better world for all: Understanding and leveraging communities in micro-lending recommendation. In *Proc. the International Conference on World Wide Web (WWW)*, pages 249–260, 2014.

[9] A. Cichocki, R. Zdunek, and S.-i. Amari. Hierarchical ALS algorithms for nonnegative matrix and 3d tensor factorization. In *Independent Component Analysis and Signal Separation*, pages 169–176. Springer, 2007.

[10] P. Dao, K. Wang, C. Collins, M. Ester, A. Lapuk, and S. C. Sahinalp. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13):i205–i213, 2011.

[11] J.-Y. Delort and E. Alfonseca. DualSum: a topic-model based approach for update summarization. In *Proc. the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223, 2012.

[12] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in Neural Information Processing Systems (NIPS)*, pages 283–290, 2005.

[13] G. Dong and J. Bailey. *Contrast Data Mining: Concepts, Algorithms, and Applications*. CRC Press, 2012.

[14] G. H. Golub and C. F. van Loan. *Matrix Computations, third edition*. Johns Hopkins University Press, Baltimore, 1996.

[15] S. K. Gupta, D. Phung, B. Adams, and S. Venkatesh. Regularized nonnegative shared subspace learning. *Data mining and knowledge discovery (DMKD)*, 26(1):57–97, 2013.

[16] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. the 22nd Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57, 1999.

[17] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[18] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[19] J. Kim, R. D. Monteiro, and H. Park. Group sparsity in nonnegative matrix factorization. In *Proc. the 2012 SIAM International Conference on Data Mining (SDM)*, pages 851–862, 2012.

[20] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[21] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[22] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 897–904. 2008.

[23] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS) 13*, pages 556–562, 2000.

[24] L. Li, G. Lebanon, and H. Park. Fast bregman divergence nmf using taylor expansion and coordinate descent. In *Proc. the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 307–315, 2012.

[25] J. Liu, C. Wang, J. Gao, and J. Han. Multi-View clustering via joint nonnegative matrix factorizations. In *Proc. the SIAM International Conference on Data Mining (SDM)*, pages 252–260, 2013.

[26] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[27] O. Odibat and C. K. Reddy. Efficient mining of discriminative co-clusters from gene expression data. *Knowledge and Information Systems*, 41(3):667–696, 2014.

[28] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons. Text mining using non-negative matrix factorizations. In *Proc. SIAM International Conference on Data Mining (SDM)*, pages 452–456, 2004.

[29] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proc. the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 650–658, 2008.

[30] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, pages 267–273, 2003.

[31] J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: Maximum margin supervised topic models for regression and classification. In *Proc. the 26th Annual International Conference on Machine Learning (ICML)*, pages 1257–1264, 2009.