2016 IEEE 16th International Conference on Data Mining

L-EnsNMF: Boosted Local Topic Discovery via Ensemble of Nonnegative Matrix Factorization

Sangho Suh
Korea University
Seoul, South Korea
sh31659@gmail.com

Jaegul Choo
Korea University
Seoul, South Korea
jchoo@korea.ac.kr

Joonseok Lee
Google Research
Mountain View, CA, USA
joonseok@google.com

Chandan K. Reddy
Virginia Tech
Arlington, VA, USA
reddy@cs.vt.edu

***Abstract*—Nonnegative matrix factorization (NMF) has been widely applied in many domains. In document analysis, it has been increasingly used in topic modeling applications, where a set of underlying topics are revealed by a low-rank factor matrix from NMF. However, it is often the case that the resulting topics give only general topic information in the data, which tends not to convey much information. To tackle this problem, we propose a novel ensemble model of nonnegative matrix factorization for discovering high-quality local topics. Our method leverages the idea of an ensemble model, which has been successful in supervised learning, into an unsupervised topic modeling context. That is, our model successively performs NMF given a residual matrix obtained from previous stages and generates a sequence of topic sets. Our algorithm for updating the input matrix has novelty in two aspects. The first lies in utilizing the residual matrix inspired by a state-of-the-art gradient boosting model, and the second stems from applying a sophisticated local weighting scheme on the given matrix to enhance the locality of topics, which in turn delivers high-quality, focused topics of interest to users. We evaluate our proposed method by comparing it against other topic modeling methods, such as a few variants of NMF and latent Dirichlet allocation, in terms of various evaluation measures representing topic coherence, diversity, coverage, computing time, and so on. We also present qualitative evaluation on the topics discovered by our method using several real-world data sets.**

***Index Terms*—Topic modeling; ensemble learning; matrix factorization; gradient boosting; local weighting.**

I. Introduction

Topic modeling has been an active area of research owing to its capability to provide a set of topics in terms of their representative keywords, which serve as a summary about large-scale document data [3]. Roughly speaking, two different topic modeling approaches exist: 1) *probabilistic models* such as probabilistic latent semantic indexing (pLSI) [15] and latent Dirichlet allocation (LDA) [3], and 2) *matrix factorization methods* such as nonnegative matrix factorization (NMF) [26].

In both types of methods, the main focus is to find a given number of bases or probability distributions, which we call *topics*, over the dictionary so that they can explain individual documents as much as possible. Because of this nature, the identified topics tend to be general ones prevalent among the entire set of documents. However, such dominant topics may not give us much meaningful information, and/or sometimes they become highly redundant with each other. This problem often arises in real-world document data when most of them share some common characteristics in their contents and/or the documents contain a large amount of noise, e.g., Twitter data.

For instance, Fig. 1 shows the sampled topics from those research papers in data mining domains[1] containing keywords 'dimension' or 'reduction.' Fig. 1(a), where standard NMF returns 'dimension' or 'reduction' as dominant keywords in most of the topics, renders the corresponding topics redundant, thus less informative.

To tackle this problem, we propose a novel topic modeling approach by building an ensemble model of NMF, which can reveal not only dominant topics but also minor but meaningful, important topics to users. Based on a gradient boosting framework, which is one of the most effective ensemble approaches, our method performs multiple stages of NMF on a residual matrix that represents the unexplained part of data from previous stages. Furthermore, we propose a novel local weighting technique combined with our ensemble method to discover diverse localized topics. As a result, unlike the highly-redundant topics of standard NMF (Fig. 1(a)), our proposed method shows much more meaningful, diverse topics, thereby allowing users to obtain deep insight, as seen in Fig. 1(b).

Overall, the main contributions of this paper are summarized as follows:

1. We develop an ensemble approach of nonnegative matrix factorization based on a gradient-boosting framework. We show that this novel approach can extract high-quality local topics from noisy documents dominated by a few uninteresting topics.

2. We perform an extensive quantitative analysis using various document datasets and demonstrate the superiority of our proposed method.

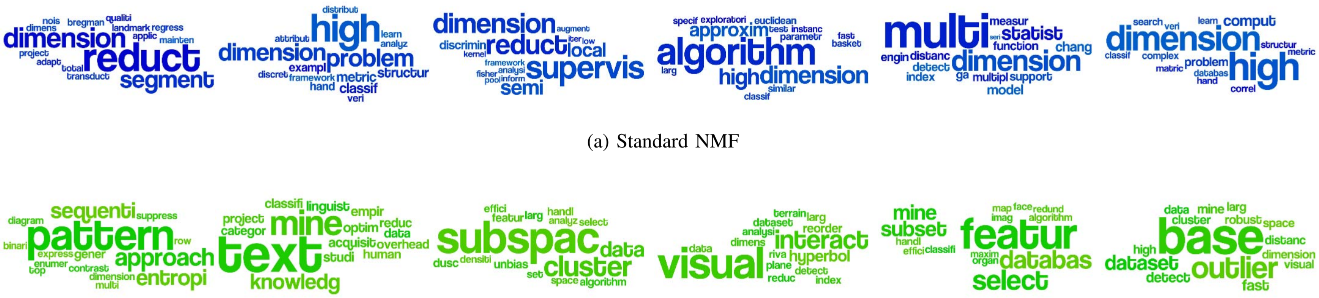
3. We show high-quality localized topic examples from several real-world datasets including research paper collections and information-scarce Twitter data.

The rest of this paper is organized as follows. Section II discusses related work. Section III describes our ensemble NMF approach, which can reveal diverse localized topics from text data. Section IV shows quantitative comparison results and

Jaegul Choo is the corresponding author.

[1] https://github.com/sanghosuh/four_area_data-matlab/

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.108



(a) Standard NMF

(b) Our approach

Fig. 1: Topic examples extracted from research papers in the data mining area published in 2000 - 2008

qualitative topic examples using various real-world datasets. Finally, Section V concludes the paper with future work.

II. Related Work

Since NMF was originally proposed by Paatero and Tapper [33] as the name of positive matrix factorization, a myriad of research about NMF has been conducted. Among them, Lee and Seung proposed the current popular form of NMF [26]. To improve the performance and the convergence properties of NMF, many studies presented an efficient alternating nonnegative least squares (ANLS)-based framework [18], [30] and its hierarchical version (HALS) [6]. In addition, Kim and Park proposed the active-set-like fast algorithms [21]. On the other hand, NMF has been applied in various manner, e.g., handling user inputs [5] and multiple data sets [16].

Related to our approach, Biggs et al. [2] proposed a successive rank-one matrix approximation based on the fact that the rank-one factorization of a nonnegative matrix has the same solution as singular value decomposition. However, their method requires to determine an optimal submatrix for such rank-one approximation, which is computationally expensive. More recently, Gillis and Glineur [11] proposed another recursive approach called nonnegative matrix underapproximation based on the additional constraints that the approximated values should be strictly smaller than the corresponding values in a given matrix, and due to this constraint, the algorithm becomes more complicated and computationally intensive compared to standard NMF. On the other hand, NMF has been used in the ensemble framework in many other machine learning applications, including clustering [13], classification [37], and bioinformatics [38].

In general, most of these existing ensemble methods primarily focus on aggregating the outputs from multiple individual models constructed independently with some variations on either an input matrix or other parameter settings. Thus, these are not applicable in topic modeling where we focus on the learned bases themselves. Furthermore, none of them has tackled the idea of constructing an ensemble of NMF models based on a gradient boosting framework, which grants a clear novelty of our work.

Without nonnegativity constraint, an ensemble of general matrix factorization has also been an active research topic in the context of collaborative filtering [35]. Ensembles of maximum margin matrix factorizations (MMMF) improved the result of a single MMMF model [7]. Ensembles of the Nystrom method [25] and of the divide-and-conquer matrix factorization [31] have also been shown effective. The Netflix Prize runner-up [34] proposed a feature-weighted least squares method using a linear ensemble of learners with human-crafted dynamic weights. Lee et al. [29] proposed a stage-wise feature induction approach, automatically inducing local features instead of human-crafted features. Local low-rank matrix factorization (LLORMA) [27], [28] combined the SVD-based matrix factorization results from locally weighted matrices under the assumption that the given matrix is only locally low-rank. It shares with our proposed method some common aspects: learning and combining locally-weighted models based on random anchor point. However, the main difference is that we impose nonnegativity in each individual model, which is more appropriate in some applications such as topic modeling. More importantly, in each stage, we systematically focus on the unexplained part of the matrix with previous ensembles, in contrast to a random choice with LLORMA.

III. L-EnsNMF

In this section, we first review standard NMF and its applications to topic modeling. Afterwards, we formulate our method called L-EnsNMF, the gradient-boosted ensemble NMF for local topic discovery,[2] as illustrated in Fig. 2.

A. Preliminaries: NMF for Topic Modeling

Given a nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$, and an integer $k \ll \min(m, n)$, nonnegative matrix factorization (NMF) [26] finds a lower-rank approximation given by

$$X \approx WH, \tag{1}$$

[2] The code is available at https://github.com/sanghosuh/lens_nmf-matlab

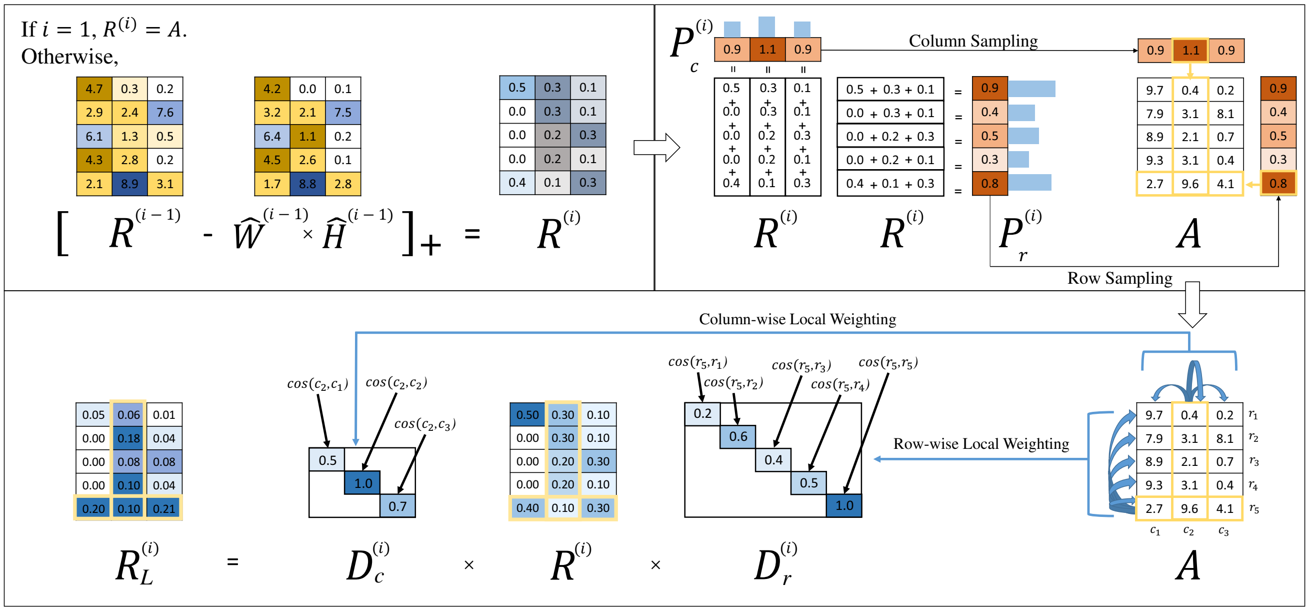


Fig. 2: Overview of the proposed ensemble approach

Notation	Description
m	Number of keywords
n	Number of documents
k_s	Number of topics per stage
p	Number of stages in L-EnsNMF
$k\ (= k_s p)$	Number of total topics
$A \in \mathbb{R}_+^{m\times n}$	Input term-by-document matrix
$\hat{W}^{(i)} \in \mathbb{R}_+^{m\times k}$	Term-by-topic matrix obtained at stage i
$\hat{H}^{(i)} \in \mathbb{R}_+^{k\times n}$	Topic-by-document matrix at stage i
$R^{(i)} \in \mathbb{R}_+^{m\times n}$	Residual matrix at stage i
$R_L^{(i)} \in \mathbb{R}_+^{m\times n}$	Localized residual matrix at stage i
$D_r^{(i)} \in \mathbb{R}_+^{m\times m}$	Row-wise scaling matrix at stage i
$D_c^{(i)} \in \mathbb{R}_+^{n\times n}$	Column-wise scaling matrix at stage i

TABLE I: Notations used in the paper

where $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{k\times n}$ are nonnegative factors. NMF is typically formulated in terms of the Frobenius norm as

$$\min_{W,\,H\geq 0} \|X - WH\|_F^2\,. \tag{2}$$

where '$\geq$' applies to every element of the given matrix in the left-hand side. In the topic modeling context, $x_i \in \mathbb{R}_+^{m\times 1}$, the i-th column of X, corresponds to the bag-of-words representation of document i with respect to m keywords, possibly with some pre-processing, e.g., inverse-document frequency weighting and column-wise ℓ_2-norm normalization. k corresponds to the number of topics. $w_l \in \mathbb{R}_+^{m\times 1}$, the l-th nonnegative column vector of W, represents the l-th topic as a weighted combination of m keywords. A large value indicates a close relationship of the topic to the corresponding keyword. The i-th column vector of H, $h_i \in \mathbb{R}_+^{k\times 1}$, represents document i as a weighted combination of k topics. Table I summarizes the notations used throughout this paper.

B. *Ensemble NMF Approach for Localized Topic Modeling*

We propose an ensemble model for topic modeling where an individual learner corresponds to NMF. Given a nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$, we learn an additive model $\hat{X}^{(q)}$ with q products $W^{(i)}H^{(i)}$:

$$X \approx \hat{X}^{(q)} = \sum_{i=1}^{q} W^{(i)}H^{(i)} \tag{3}$$

where $W^{(i)} \in \mathbb{R}_+^{m\times k_s}$, $H^{(i)} \in \mathbb{R}_+^{k_s\times n}$ and q is the number of individual learners. That is, the i-th *stage* represents a local NMF model discovering the i-th k_s local topics. To achieve this approximation, we introduce an objective function in terms of the Frobenius norm as follows:

$$\min_{W^{(i)},H^{(i)}\geq 0,\, i=1,\cdots,q} \left\| X - \sum_{i=1}^{q} W^{(i)}H^{(i)} \right\|_F^2 . \tag{4}$$

Our proposed method solves this problem in a forward stage-wise manner [14], inspired by well-known ensemble learning methods in a supervised learning context such as AdaBoost [9] and gradient boosting [10]. We iteratively add a new local model to better approximate X, fitting the i-th local NMF, $W^{(i)}H^{(i)}$, with rank k_s to the localized residual, which is the unexplained portion by previously learned $i-1$ local models. To this end, let us first define the (non-localized) nonnegative residual matrix at stage i as

$$R^{(i)} = \begin{cases} X & \text{if} \quad i = 1 \\ \left[R^{(i-1)} - W^{(i-1)}H^{(i-1)}\right]_+ & \text{if} \quad i \geq 2 \end{cases} \tag{5}$$

where $[\cdot]_+$ is an operator that converts every negative element in the matrix to zero. Next, we apply local weighting on this residual matrix $R^{(i)}$ to obtain its localized version $R_L^{(i)}$ and compute $W^{(i)}$ and $H^{(i)}$ by applying NMF to $R_L^{(i)}$ as an input matrix. More details about our local weighting scheme will be described in Section III-E.

In general, the input matrix to NMF at stage i is defined as

$$\begin{aligned} R^{(i)} &= \left[\left[\left[X - W^{(1)}H^{(1)}\right]_+ - W^{(2)}H^{(2)}\right]_+ \cdots \right. \\ &\qquad \left. - W^{(i-1)}H^{(i-1)}\right]_+, \end{aligned} \tag{6}$$

where $\hat{W}^{(i)}$ and $\hat{H}^{(i)}$ are obtained in a forward stage-wise manner, e.g., $\left(\hat{W}^{(1)}, \hat{H}^{(1)}\right)$, $\left(\hat{W}^{(2)}, \hat{H}^{(2)}\right)$, and so on. By a simple manipulation, one can prove that our original objective function shown in Eq. (4) is equivalent to a single-stage NMF as

$$\min_{W^{(i)}, H^{(i)} \geq 0,\, i=1,\ldots,q} \left\| X - \sum_{i=1}^{q} W^{(i)}H^{(i)} \right\|_F^2 \tag{7}$$

$$= \min_{W^{(i)}, H^{(i)} \geq 0,\, i=1,\ldots,q} \| X - WH \|_F^2 \tag{8}$$

where $W = \begin{bmatrix} W^{(1)} & W^{(2)} & \cdots & W^{(q)} \end{bmatrix} \in \mathbb{R}_+^{m \times (k_s q)}$ and $H = \begin{bmatrix} H^{(1)} \\ H^{(2)} \\ \vdots \\ H^{(q)} \end{bmatrix} \in \mathbb{R}_+^{(k_s q) \times n}$.

However, the main difference between our method and the (single-stage) standard NMF lies in the approach adopted to solve W (or $W^{(i)}$'s) and H (or $H^{(i)}$'s). That is, in standard NMF, all of $W^{(i)}$'s and $H^{(i)}$'s are optimized simultaneously within a single optimization framework using various algorithms such as a gradient descent [30], a coordinate [26], or a block-coordinate descent framework [19]. However, our proposed method solves each set of $(W^{(i)}, H^{(i)})$'s in a greedy, sequential manner, which means that once the solution for $(W^{(i)}, H^{(i)})$ is obtained at stage i, it is fixed during the remaining iterations.

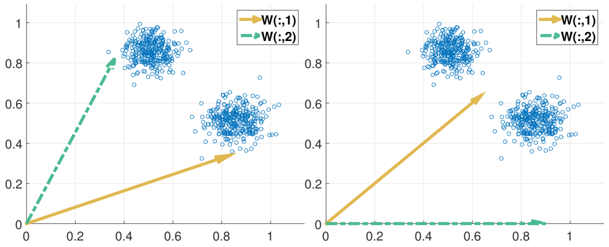
Our approach can be viewed as a functional gradient boosting approach [14]. In detail, let $f^{(i)}$ and L be

$$f^{(i)} = f\left(W^{(1)}, \cdots, W^{(i)}, H^{(1)}, \cdots, H^{(i)}\right) = \sum_{l=1}^{i} W^{(l)}H^{(l)},$$

$$L\left(X, f^{(i)}\right) = \left\| X - f^{(i)} \right\|_F^2 = \left\| X - \sum_{l=1}^{i} W^{(l)}H^{(l)} \right\|_F^2, \tag{9}$$

respectively. In the case where $f^{(i)} = f^{(i-1)}$, which corresponds to the results from the previous stage $i-1$, the gradient of Eq. (9), $\mathbf{g}_i$, can be expressed as

$$\begin{aligned} \mathbf{g}_i &= \left[\frac{\partial L\left(X, f^{(i)}\right)}{\partial f^{(i)}} \right]_{f^{(i)} = f^{(i-1)}} \\ &= 2\left(X - f^{(i-1)}\right) = 2\left(X - \sum_{l=1}^{i-1} W^{(l)}H^{(l)}\right). \end{aligned}$$



(a) Standard NMF

(b) Deflation-based NMF

Fig. 3: Synthetic data example where $m = 2$, $k_s = 1$, and $q = 2$

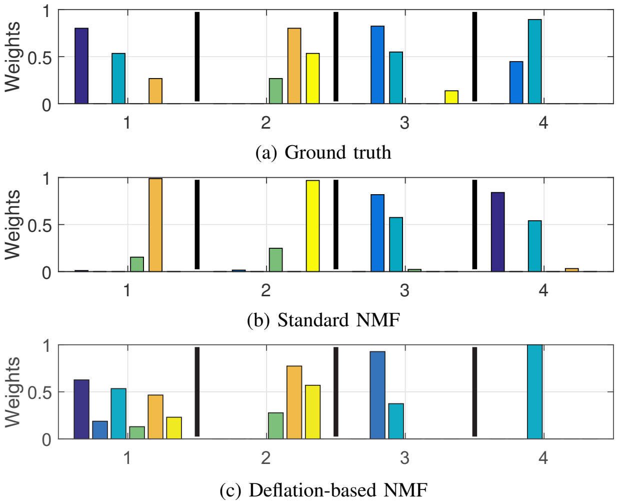
Now, imposing the constraints $f^{(i)} \geq 0$ due to $W^{(i)}, H^{(i)} \geq 0$ and ignoring the constant in the above equation, we can obtain the projected gradient $P\left[\mathbf{g}_i\right]$ as Eq. (6) by setting $i = 1, \cdots, q$.

C. Why NMF on Residual Matrices

Traditionally, a greedy approach such as the one we proposed in Section III-B can be viewed as a rank-deflation procedure for low-rank matrix factorization, which obtains low-rank factors one at a time [36]. The power method [12], which consecutively reveals the most dominant eigenvalue and vector pairs, is a representative deflation method. It is known that the solution obtained by such a (greedy) deflation procedure is equivalent to the solution obtained by simultaneously optimizing all the low-rank factors in singular value decomposition [12], where the low-rank factor matrices are allowed to be both positive and negative.

Generally, such a deflation method does not work for NMF, due to the limitation that the factor matrices should not contain negative elements. Fig. 3 shows the comparison between standard NMF and our ensemble approach, given a synthetic Gaussian mixture data in a two-dimensional feature space. As seen in Fig. 3(a), the column vectors of W generated from standard NMF in Eq. (2) successfully reveal the two components of the Gaussian mixture data. However, in the deflation approach shown in Fig. 3(b), the basis vector at the first stage, $W^{(1)} \in \mathbb{R}_+^{2 \times 1}$, is computed as a global centroid and then at the second stage, $W^{(2)} \in \mathbb{R}_+^{2 \times 1}$, which is computed on the residual matrix, is shown as the vector along a single axis, y-axis in this case. As a result, the two bases found by the deflation-based NMF approach fail to identify the true bases. This is clearly the case where the deflation approach does not work with NMF.

In the case of text data, however, where the dimension is high and the matrix is highly sparse, we claim that such a deflation method can work as well as or even better than standard NMF. Fig. 4 shows another synthetic data example where the data are relatively high-dimensional compared to those in the previous example, e.g., $m = 5$, and the column vectors of the true W are sparse. We generated synthetic data using a Gaussian mixture with the mean values of its components equal to the columns of W shown in Fig. 4(a). In this figure, standard NMF (Fig. 4(b)) does not properly



(a) Ground truth

(b) Standard NMF

(c) Deflation-based NMF

Fig. 4: Column vectors of W from synthetic data with $m = 5$, $k_s = 1$, and $q = 4$. The columns of W's generated by both the standard and the ensemble NMF have been aligned to those of the ground truth W using the Hungarian method [24].

recover the true column vectors of W except for the third component. On the other hand, our deflation-based NMF approach (Fig. 4(c)) recovers most of the true column vectors of W much better than the standard NMF.

The reason why the deflation-based NMF works surprisingly well with sparse high-dimensional data, e.g., text data, is because their original dimensions, e.g., keywords in text data, with large values are unlikely to overlap among different column vectors of W due to its sparsity. In this case, the deflation-based NMF can be suitable by finding these dimensions or keywords with large values in one vector at a time. Combined with our local weighting technique described in Section III-E, such a deflation-based method helps to reveal highly non-redundant, diverse topics from the data by preventing the significant keyword shown in a particular topic from appearing in the other topics.

D. Efficient Algorithm for Ensemble NMF

A unique advantage of our method is that regardless of the total number of topics, k, one can keep the rank used in computing NMF at each stage, k_s, small while increasing the number of stages, q, i.e., $k_s \ll (k = k_s q)$. Hence, to efficiently solve NMF with a low value of k_s, we extend a recent active-set-based NMF algorithm [23], which demonstrated significantly high efficiency for a small value of k_s.

In detail, our algorithm is built upon the two-block coordinate descent framework, which iteratively solves W while fixing H and then the other way around. Given a local residual matrix $R_L^{(i)}$ at stage i, we first obtain the term-by-topic matrix $\hat{W}^{(i)}$ and the topic-by-document matrix $\hat{H}^{(i)}$ by solving

$$\left(W^{(i)},\, H^{(i)}\right) = \underset{W,H\geq 0}{\arg\min} \left\| R_L^{(i)} - WH \right\|_F^2 . \tag{10}$$

Each sub-problem of solving $W^{(i)}$ and $H^{(i)}$ in the above equation can be represented as

$$\min_{G\geq 0} \|Y - BG\|_F^2 = \sum_i \min_{\mathbf{g}_i \geq 0} \|\mathbf{y}_i - B\mathbf{g}_i\|_2^2 \tag{11}$$

where H is obtained by setting $B = W$, $G = H$, and $Y = X$, W is obtained by setting $B = H$, $G = W$, and $Y = X^T$, and $\mathbf{g}_i$ and $\mathbf{y}_i$ are the i-th columns of G and Y, respectively. Let us consider each problem in the summation operator and rewrite it as

$$\min_{\mathbf{g}\geq 0} \|\mathbf{y} - B\mathbf{g}\|_2^2 , \tag{12}$$

which is a nonnegativity-constrained least squares problem. Here, the elements of the vector $\mathbf{g}$ can be partitioned into the one containing zeros and the other containing strictly positive values, and let us call these sets of dimension indices of the active and the passive sets as $\mathcal{I}_a$ and $\mathcal{I}_p$, respectively. Once we fully know $\mathcal{I}_a$ and $\mathcal{I}_p$ for the optimal solution of Eq. (12), such an optimal solution is equivalent to the solution obtained by solving an unconstrained least squares using only the passive set of variables [20], i.e.,

$$\min \|B\left(:, \mathcal{I}_p\right) \mathbf{g}_i \left(\mathcal{I}_p\right) - \mathbf{y}\|_2^2 . \tag{13}$$

The active-set method iteratively modifies the partitioning between $\mathcal{I}_a$ and $\mathcal{I}_p$ and solves for Eq. (13) until the optimal $\mathcal{I}_a$ and $\mathcal{I}_p$ are found. However, this process is performed one at a time for a particular partitioning until convergence, which requires a large number of iterations. The approach proposed in [23] accelerates this process for small k_s values by exhaustively solving based on all the possible partitionings and selecting the optimal one since the number of all the different partitionings, which is 2^{k_s}, would remain small.

However, this approach is not applicable when k_s is large since the number of partitionings grows exponentially with respect to k_s, and thus the original approach [23] proposed to build a hierarchical tree until the method obtains the number of leaf nodes as the total number of clusters or topics. However, in this paper, we adopt this exhaustive search approach for an optimal active/passive set partitioning as our individual learner at each stage, which maintains the small value of k_s when solving NMF at each stage. As will be shown in Section IV, our method does not only generate high-quality local topics but also provides high computational efficiency compared to standard NMF for obtaining the same number of topics.

E. Local Weighting

In contrast to standard NMF, which discovers mostly general but uninteresting topics, our ensemble approach tends to identify major but uninteresting topics at an early stage and gradually reveal interesting local topics in subsequent stages, since minor, unexplained topics will become more prominent in the residual matrix as stages proceed. However, when the number of topics per stage k_s is small, we found that this process sometimes takes many stages before revealing interesting topics. To further accelerate this process and enhance the diversity of local topics, we perform local weighting on the residual matrix $R^{(i)}$ so that the explained parts are suppressed while the unexplained parts are highlighted.

We form the localized residual matrix $R_L^{(i)}$ as

$$R_L^{(i)} = D_r^{(i)} R^{(i)} D_c^{(i)}, \tag{14}$$

where diagonal matrices $D_r^{(i)} \in \mathbb{R}_+^{m\times m}$ and $D_c^{(i)} \in \mathbb{R}_+^{n\times n}$ perform row- and column-wise scaling, respectively. Solving NMF given this scaled residual matrix is equivalent to solving a weighted version of NMF with the corresponding row- and column-wise scaling since

$$\begin{aligned} & \min_{W^{(i)},H^{(i)}\geq 0} \left\| D_r^{(i)} \left(R^{(i)} - W^{(i)} H^{(i)} \right) D_c^{(i)} \right\|_F^2 \\ = & \min_{W^{(i)},H^{(i)}\geq 0} \left\| D_r^{(i)} R^{(i)} D_c^{(i)} - D_r^{(i)} W^{(i)} H^{(i)} D_c^{(i)} \right\|_F^2 \\ = & \min_{W_L^{(i)},H_L^{(i)}\geq 0} \left\| R_L^{(i)} - W_L^{(i)} H_L^{(i)} \right\|_F^2 \end{aligned}$$

by setting $W_L^{(i)} = D_r^{(i)} W^{(i)}$ and $H_L^{(i)} = H^{(i)} D_c^{(i)}$.

We design these scaling factors to assign higher weights to those rows or columns less explained (large residuals) by previous stages. Let us define the probability distributions $P_r^{(i)}$ and $P_c^{(i)}$ over row indices, x's, and over column indices, y's, respectively, as

$$P_r^{(i)}(x) = \frac{\sum_{s=1}^{n} R^{(i)}(x,s)}{\sum_{l=1}^{m}\sum_{s=1}^{n} R^{(i)}(l,s)} \text{ for } x = 1,\cdots,m \tag{15}$$

$$P_c^{(i)}(y) = \frac{\sum_{l=1}^{m} R^{(i)}(l,y)}{\sum_{l=1}^{m}\sum_{s=1}^{n} R^{(i)}(l,s)} \text{ for } y = 1,\cdots,n. \tag{16}$$

In Eqs. (15) and (16), higher probability values are assigned to those rows or columns with larger values in residual matrix $R^{(i)}$. In other words, a higher probability indicates that the corresponding row or column is less explained up to the previous stage. Rather than directly using these probability distributions as the local weighting matrices $D_r^{(i)}$ or $D_c^{(i)}$, we sample from this probability distribution only a single row a_r and a column a_c, which we call an *anchor point*, corresponding to a particular keyword and a document that were not yet well explained from previous stages, respectively. The purpose of this selection process is to allow the NMF computation with only a small k_s to properly reveal the topics around the selected document and keyword, rather than to generate still unclear topics reflecting most of the unexplained documents.

The diagonal entries of $D_r^{(i)}$ and $D_c^{(i)}$ are then computed based on the similarity of each row and column to the anchor row a_r and column a_c, respectively. Specifically, given the selected a_r and a_c, we use the cosine similarity to compute the l-th diagonal entry of $D_r^{(i)}(l,l)$ and the s-th diagonal entry of $D_c^{(i)}(s,s)$, respectively, as

$$D_r^{(i)}(l,l) = \cos\left(X(a_r,:), X(l,:)\right) \text{ for } l = 1,\cdots,m \tag{17}$$

$$D_c^{(i)}(s,s) = \cos\left(X(:,a_c), X(:,s)\right) \text{ for } s = 1,\cdots,n. \tag{18}$$

Using these weights, we enhance the locality of the resulting topics.

Applying the localized residual matrix as described above, we plug $R_L^{(i)}$ (Eq. (14)) into Eq. (10) and obtain $W^{(i)}$ and $H^{(i)}$. When computing the residual matrix in the next stage using $W^{(i)}$ and $H^{(i)}$, as shown in Eq. (5), however, it may end up removing only the fraction of the residuals, which can be significantly smaller than the unweighted residuals since all the weights are less than or equal to 1. To adjust this shrinking effect caused by local weighting, we recompute $H^{(i)}$ using the given $W^{(i)}$ and the non-weighted residual matrix $R^{(i)}$, i.e.,

$$H^{(i)} = \arg\min_{H\geq 0} \left\| W^{(i)} H - R^{(i)} \right\|_F^2. \tag{19}$$

In this manner, our approach still maintains the localized topics $W^{(i)}$ from $R_L^{(i)}$ while properly subtracting the full portions explained by these topics from $R^{(i)}$ for the next stage.

Finally, the detailed algorithm of our approach is summarized in Algorithm 1.

Algorithm 1: Localized Ensemble NMF (**L-EnsNMF**)

```
Input: Input matrix $X \in \mathbb{R}_+^{m\times n}$, integers $k_s$ and $q$
Output: $W^{(i)} \in \mathbb{R}_+^{m\times k_s}$ and $H^{(i)} \in \mathbb{R}_+^{k_s\times n}$ for
        $i = 1,\cdots,q$
for $i = 1$ to $q$ do
    Compute $R^{(i)}$ using Eq. (6).
    Compute $P_r^{(i)}(x)$ and $P_c^{(i)}(y)$ using Eqs. (15)
      and (16).
    $a_r \leftarrow$ Sample a row from $P_r^{(i)}(x)$.
    $a_c \leftarrow$ Sample a column from $P_c^{(i)}(y)$.
    Compute $D_r^{(i)}$ and $D_c^{(i)}$ using Eqs. (17) and (18).
    Compute $R_L^{(i)}$ using Eq. (14).
    Compute $W^{(i)}$ using Eq. (10).
    Compute $H^{(i)}$ using Eq. (19).
end
```

IV. Experiments

In this section, we present extensive quantitative comparisons of our proposed approach against other state-of-the-art methods. Afterwards, we demonstrate qualitative results containing high-quality localized topics identified by our methods, which would be otherwise difficult to discover using other existing methods, from several real-world datasets.

All the experiments were conducted using MATLAB 8.5 (R2015a) on a desktop computer with dual Intel Xeon E5-2687W processors.

A. Experimental Setup

In the following, we describe our experimental setup including datasets, baseline methods, and evaluation measures.

1) Datasets: We selected the following five real-world document datasets: 1) Reuters-21578 (**Reuters**),[3] a collection of articles from the Reuters newswire in 1987; 2) 20 Newsgroups (**20News**),[4] from Usenet newsgroups; 3) **Enron**[5] containing 2,000 randomly sampled emails generated by the employees of Enron Corporation; 4) IEEE-Vis (**VisPub**),[6] academic papers published in IEEE Visualization conferences (SciVis, InfoVis, and VAST) from 1990 to 2014; and 5) **Twitter**, a collection of 2,000 randomly selected tweets generated from a specific location of New York City in June 2013. These datasets are summarized in Table II.

	Reuters	20News	Enron	VisPub	Twitter
#docs	7,984	18,221	2,000	2,592	2,000
#words	12,411	36,568	19,589	7,535	4,212

TABLE II: Summary of the data sets used

2) Baseline Methods : We compared our method, L-EnsNMF, against various state-of-the-art methods, including standard NMF (**StdNMF**) [19],[7] sparse NMF (**SprsNMF**) [17],[8] orthogonal NMF (**OrthNMF**) [8],[9] and latent Dirichlet allocation (**LDA**) [3].[10]

In most of these methods, we used default parameter values provided by the software library, including the regularization parameters for SprsNMF, OrthNMF, and LDA, as well as the parameters used in convergence criteria. Since there exist no clear convergence criteria for the Gibbs sampling-based implementation of LDA, we set the number of iterations as 2,000, which is one of the most common settings. Also, note that we did not use LLORMA as one of the baseline methods because it is a supervised method and does not impose a nonnegativity constraint, the two characteristics of which make it unfit for topic modeling.

3) Evaluation Measures: We adopted several evaluation measures for assessing the quality of the generated topics: topic coherence [1] and the total document coverage. Additionally, we compared the computing times between different methods. In the following, we will describe each measure in detail.

Topic Coherence. To assess the quality of individual topics, we utilize the point-wise mutual information (PMI) [32], which indicates how likely a pair of keywords co-occur in the same document. That is, given two words w_i and w_j, PMI is defined as

$$\text{PMI}\left(w_i, w_j\right) = \log \frac{p\left(w_i, w_j\right)}{p\left(w_i\right) p\left(w_j\right)}, \tag{20}$$

where $p\left(w_i, w_j\right)$ represents the probability of w_i and w_j co-occurring in the same document and $p\left(w_i\right)$ represents the probability of w_i occurring in our document data set. Thus, a pair of words with a high PMI score can be viewed as being semantically related, thus conveying meaningful information. To extend this notion at a topic level and compute the topic coherence measure, we first select the ten most representative keywords of each topic and then compute the average PMI score among them. Next, we further compute the average of this score over all the given topics.

TABLE III: Comparison of topic coherence values. The reported results are averaged values over 20 runs. The best performance values are shown in **bold**, and the second best ones are underlined.

	k = 12 (k_s = 2, q = 6)				
	Std NMF	Sprs NMF	Orth NMF	LDA	L-Ens NMF
Reuters	1.051	1.121	0.631	**1.348**	1.315
20News	1.435	1.537	0.920	1.685	**2.108**
Enron	1.918	1.980	1.885	1.778	**2.490**
VisPub	0.615	0.562	0.619	0.367	**0.769**
Twitter	1.426	1.649	1.431	0.487	**2.761**
	k = 24 (k_s = 2, q = 12)				
	Std NMF	Sprs NMF	Orth NMF	LDA	L-Ens NMF
Reuters	1.213	1.408	0.874	1.399	**1.640**
20News	1.512	1.795	1.000	2.043	**2.334**
Enron	1.890	1.792	1.886	1.928	**2.370**
VisPub	0.645	0.358	0.645	0.548	**0.940**
Twitter	1.654	1.764	1.671	0.442	**2.843**
	k = 48 (k_s = 2, q = 24)				
	Std NMF	Sprs NMF	Orth NMF	LDA	L-Ens NMF
Reuters	1.349	1.322	1.103	1.590	**1.832**
20News	1.637	1.864	1.086	2.180	**2.375**
Enron	1.839	1.881	1.841	2.065	**2.327**
VisPub	0.737	0.918	0.745	0.648	**1.136**
Twitter	1.591	1.488	1.731	0.439	**2.958**

Total Document Coverage. This measure computes how many documents (out of the entire document set) can be explained by a given set of topics. Here, a document is said to be *explained* if there exists a topic such that at least a certain number of keywords among its most representative keywords are found in that document. That is, given a set of topics $\mathcal{T} \in \{t_1, \cdots, t_k\}$ and a set of documents $\mathcal{D} = \{d_1, \cdots, d_n\}$, the total document coverage is defined as

$$\begin{aligned} &\text{TDC}\left(\mathcal{T}, \mathcal{D}\right) \\ &= \frac{\left|d \in \mathcal{D} : \exists t_i \in \mathcal{T} \text{ s.t. } \left|w\left(d\right) \cap w_R\left(t_i, c_1\right)\right| \geq c_2\right|}{\left|\mathcal{D}\right|}, \end{aligned} \tag{21}$$

where $w\left(d\right)$ represents the set of words occurring in document d and $w_R\left(t_i, c_1\right)$ represents the set of the c_1 most representative keywords of topic t_i. In other words, this measures the relative number of documents containing at least c_2 keywords among the c_1 most representative keywords of one topic or more. In our experiment, we set $c_1 = 20$ and observed how this measure changes while varying c_2.

In terms of the comparison between two topic sets with an equal number of topics, if one set has a better value of this measure than the other, then one can view it as having not

[3] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

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

[5] https://www.cs.cmu.edu/~./enron/

[6] http://www.vispubdata.org/site/vispubdata/

[7] https://github.com/kimjingu/nonnegfac-matlab

[8] http://www.cc.gatech.edu/~hpark/software/nmf_bpas.zip

[9] http://davian.korea.ac.kr/myfiles/list/Codes/orthonmf.zip

[10] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

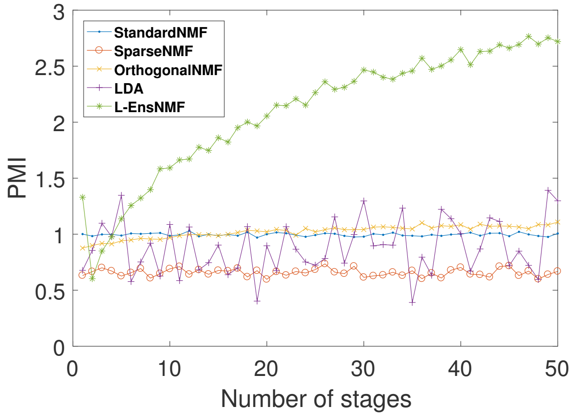


Fig. 5: Topic coherence values over stages when 100 topics (k_s = 2, q = 50) are computed. Each value of our method represents the average topic coherence value of k_s corresponding topics per stage. The results of the other methods show the average values per k_s topics. The results were obtained by computing the average values over 1,000 runs.

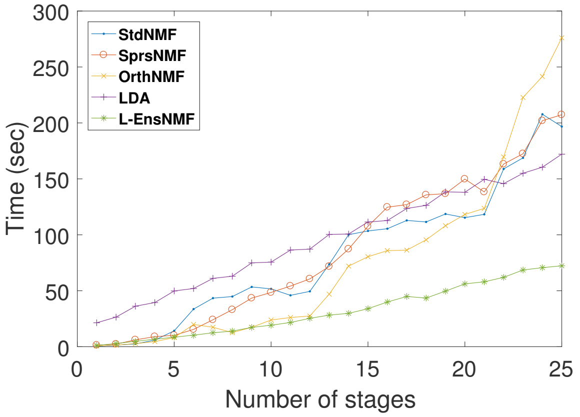


Fig. 6: Comparison of computing times for VisPub dataset. The results were obtained from the average values over 50 runs.

only the better quality of topics but also the better diversity since it explains more number of documents using the same number of topics.

B. Quantitative Analysis

Topic Coherence. Table III compares the quality of the topics generated by different topic modeling methods using the topic coherence measure. As seen in this table, our localized ensemble NMF is shown to maintain the highest topic coherence consistently in most of the cases. For Reuters dataset, with $k = 12$, LDA performs the best while our method trails behind closely with the second best coherence scores. Except for this case, however, our method demonstrates the highest performance consistently in all the datasets and the different number of topics. Note also that there is no clear second best performing method. This observation lends further support for our localized ensemble NMF by indicating that other comparable methods showing equal or even better performances at times may not perform consistently in all the datasets.

In addition, Fig. 5 shows how the topic coherence value changes as the stage proceeds in our ensemble model. Here, one can see that the topic coherence is constantly improved as the stages proceed, which ends up generating those topics with much better quality than any other methods. This strongly supports our claim that the gradient boosting-based ensemble framework for NMF works surprisingly well in topic modeling applications and that the topics generated in later stages in this framework will have significant advantages than those generated by other existing methods.

Total Document Coverage. Table IV shows the total document coverage results of different methods. In this table, our method is shown to be the best or the second best method for all the different number of topics.

Another important observation is that the performance margin between our method and the others becomes larger in favor of ours when c_2 in Eq. (21) increases. Note that a large c_2 imposes a strict condition for a particular document to be explained by a topic (Section IV-A3). The fact that our method works well compared to other methods in such a strict condition signifies its important advantage of revealing the faithful semantic information from the resulting topics.

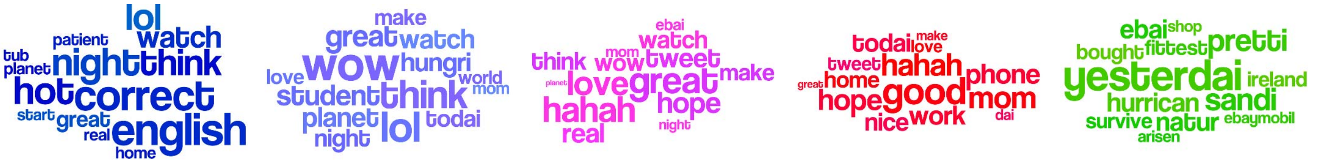
Computing Times. We measured the running time of different methods by varying the total number of topics, k, from 2 to 50. In the case of our ensemble NMF method, we fixed k_s as 2 while changing q from 1 to 25. As shown in Fig. 6, our method runs fastest, and more importantly, it scales better than any other methods with respect to k. As discussed in Section III-D, such a computational advantage is due to two synergetic aspects: (1) maintaining k_s to be small regardless of how large k is and (2) using a highly efficient NMF algorithm that performs an exhaustive search on all the possible active/passive set partitionings. Such promising aspects of our proposed L-ensNMF imply that it can be used to efficiently compute a large number of topics from large-scale data.

C. Exploratory Topic Discovery

In this section, we present diverse interesting topics uniquely found by our methods from several datasets. Fig. 7 shows the five representative topics extracted from Twitter dataset by the baseline methods and our method. The keywords found by other methods are not informative in a sense that they are either too general or common words with no interesting implication–see words, such as 'lol,' 'wow,' 'great,' 'hahah.' On the contrary, our localized ensemble NMF generates interesting keywords for its topics, e.g., 'hurricane,' 'sandi,' 'fittest,' 'survive,' 'ireland,' which deliver more specific and insightful information to users. For example, it discovered 'hurricane sandi'–which devastated New York City in 2012–while both

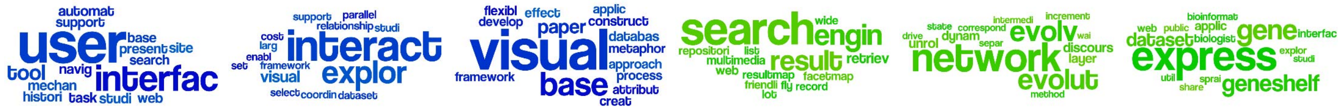
TABLE IV: Total document coverage of VisPub based on six different methods, as defined in Eq. (21). The reported results are averaged values over 20 runs. The best performance values are shown in **bold**, and the second best ones are underlined.

	c_2 in Eq. (21)	3	4	5	6	7	8	9	10	Average
k = 10 (k_s=2, q=5)	StdNMF	0.937	0.778	0.496	0.236	0.081	0.021	0.004	0.000	0.319
	SprsNMF	0.923	0.746	0.473	0.229	0.083	0.021	0.004	0.000	0.301
	OrthNMF	0.940	0.790	0.519	0.256	0.091	0.024	0.005	0.000	0.328
	LDA	**0.970**	**0.884**	**0.666**	**0.352**	0.141	0.037	0.005	0.000	**0.382**
	L-EnsNMF	0.941	0.821	0.601	0.350	**0.153**	**0.047**	**0.009**	**0.001**	0.365
k = 50 (k_s=2, q=25)	StdNMF	0.962	0.770	0.428	0.155	0.039	0.007	0.001	0.000	0.295
	SprsNMF	0.951	0.717	0.367	0.125	0.030	0.006	0.001	0.000	0.275
	OrthNMF	0.963	0.772	0.435	0.158	0.040	0.007	0.001	0.000	0.297
	LDA	**0.977**	**0.902**	0.651	0.336	0.107	0.028	0.001	0.000	0.375
	L-EnsNMF	0.972	0.892	**0.689**	**0.412**	**0.178**	**0.057**	**0.012**	**0.003**	**0.402**
k = 100 (k_s=2, q=50)	StdNMF	0.962	0.724	0.346	0.111	0.028	0.007	0.002	0.000	0.273
	SprsNMF	0.948	0.676	0.303	0.099	0.024	0.005	0.001	0.000	0.257
	OrthNMF	0.962	0.722	0.345	0.111	0.028	0.007	0.001	0.000	0.272
	LDA	0.979	**0.919**	**0.676**	0.336	0.105	0.024	0.003	0.000	0.380
	L-EnsNMF	**0.980**	0.889	0.669	**0.397**	**0.179**	**0.060**	**0.017**	**0.005**	**0.400**


lol watch patient tub planet night think hot correct start great real english home
make great watch love wow hungri world mom student think planet todai night lol
ebai watch mom wow tweet think love great make planet hahah hope real night
todai make love tweet hahah great home phone hope good mom nice work dai
ebai shop bought fittest pretti yesterdai ireland hurrican sandi survive natur ebaymobil arisen

(a) Standard NMF (b) Sparse NMF (c) Orthogonal NMF (d) LDA (e) L-Ens NMF

Fig. 7: Topic examples from Twitter dataset


automat support user base present site search tool navig mechan histori task studi web interfac
support parallel relationship studi cost larg enabl set framework interact visual explor select coordin dataset
flexibl effect develop applic paper construct visual databas metaphor approach framework base process attribut creat
search wide engin repositori list multimedia result retriev web resultmap facetmap friendli fly record lot
intermedi increment state correspond drive unrol dynam separ evolv wai network discours layer evolut method
bioinformat web public applic dataset biologist gene interfac express explor studi util sprai share geneshelf

(a) Standard NMF (b) L-EnsNMF

Fig. 8: Discovered topics using VisPub dataset

words were not found in any of the 100 topics (10 keywords each) generated by other baseline methods. This demonstrates that our method could be used in, say, early disaster detection and many other areas that can greatly benefit from local topic discovery. Besides, a quick search for related web documents with the query 'ireland hurricane sandy' led to the discovery of the local news that the Ireland football team visited New York in June 2013 to boost a community hit by Hurricane Sandy. This was another example indicative of how local topics can be more useful than global topics.

The second set of examples for assessing the semantic topic quality are extracted from VisPub dataset, as shown in Fig. 8. The results from standard NMF (Fig. 8(a)) are mostly dominated by those keywords too obvious and thus uninformative, e.g., 'visual,' 'user,' 'interface,' 'tool,' 'interact,' considering that the documents are mainly about interactive visualization and user interfaces. On the other hand, our method delivers more focused keywords revealing the useful information about specific sub-areas in the field. For example, from the topic containing 'search,' 'engine,' 'result,' and 'multimedia,' which are about search engine visualization, we found the paper "Visualizing the results of multimedia web search engines" by Mukherjea et al. The keywords, 'network' and 'evolut,' which are about dynamic, time-evolving network visualization, led us to related papers, e.g., "Visual unrolling of network evolution and the analysis of dynamic discourse" by Brandes et al. Finally, the keywords, 'gene' and 'express,' which are about biological data visualization, point directly to the paper "MulteeSum: a tool for comparative spatial and temporal gene expression data" by Meyer et al.

V. Conclusion

In this paper, we presented a novel ensemble approach of NMF for high-quality local topic discovery via a gradient boosting framework and a systematic local weighting technique. The proposed method is especially useful in disclosing local topics that are otherwise left undiscovered when using existing topic modeling algorithms. Although the algorithm is designed to find localized topics, our ensemble approach achieves outstanding performances in both topic coherence and document coverage compared to other approaches that mostly reveal general topics. This indicates that our approach does not only excel in providing meaningful topics but also represents or summarizes the overall information of a corpus better than other state-of-the art methods. Moreover, it performs much faster than other methods owing to the exhaustive search approach for an optimal active/passive set partitioning, which makes our method promising for large-scale and real-time topic modeling applications.

As our future work, we plan to expand our work to an interactive topic discovery system [4], [22] by flexibly steering the local weighting process in a user-driven manner so that the subsequent topics can properly reflect a user's subjective interest and task goals.

Acknowledgments

This work was partially supported by National Science Foundation grants IIS-1231742, IIS-1527827, and IIS-1646881 and by Institute for Information & communications Technology Promotion (IITP) and Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2015K2A1A2070536). Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of funding agencies.

References

[1] N. Aletras and M. Stevenson. Evaluating topic coherence using distributional semantics. In *Proc. the International Conference on Computational Semantics*, pages 13–22, 2013.

[2] M. Biggs, A. Ghodsi, and S. Vavasis. Nonnegative matrix factorization via rank-one downdate. In *Proc. the International Conference on Machine Learning (ICML)*, pages 64–71, 2008.

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

[4] 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.

[5] J. Choo, C. Lee, C. K. Reddy, and H. Park. Weakly supervised nonnegative matrix factorization for user-driven clustering. *Data Mining and Knowledge Discovery (DMKD)*, 29(6):1598–1621, 2015.

[6] 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. 2007.

[7] D. DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *Proc. the International Conference on Machine Learning (ICML)*, pages 249–256, 2006.

[8] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[9] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[10] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[11] N. Gillis and F. Glineur. Using underapproximations for sparse nonnegative matrix factorization. *Pattern Recognition*, 43(4):1676–1687, 2010.

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

[13] D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble non-negative matrix factorization methods for clustering protein–protein interactions. *Bioinformatics*, 24(15):1722–1728, 2008.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

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

[16] H. Kim, J. Choo, J. Kim, C. K. Reddy, and H. Park. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 567–576, 2015.

[17] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[18] 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.

[19] 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.

[20] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. 2008.

[21] 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.

[22] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2017.

[23] D. Kuang and H. Park. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 739–747, 2013.

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

[25] S. Kumar, M. Mohri, and A. Talwalkar. Ensemble nystrom method. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1060–1068, 2009.

[26] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[27] J. Lee, S. Kim, G. Lebanon, and Y. Singer. Local low-rank matrix approximation. In *Proc. the International Conference on Machine Learning (ICML)*, pages 82–90, 2013.

[28] J. Lee, S. Kim, G. Lebanon, Y. Singer, and S. Bengio. Llorma: local low-rank matrix approximation. *Journal of Machine Learning Research (JMLR)*, 17(15):1–24, 2016.

[29] J. Lee, M. Sun, S. Kim, and G. Lebanon. Automatic feature induction for stagewise collaborative filtering. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[30] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[31] L. W. Mackey, A. S. Talwalkar, and M. I. Jordan. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1134–1142, 2011.

[32] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Proc. the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 100–108, 2010.

[33] P. Paatero and U. Tapper. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[34] J. Sill, G. Takacs, L. Mackey, and D. Lin. Feature-weighted linear stacking. *Arxiv preprint 0911.0460*, 2009.

[35] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4:2–4:2, 2009.

[36] J. H. Wilkinson, J. H. Wilkinson, and J. H. Wilkinson. *The algebraic eigenvalue problem*, volume 87. Clarendon Press Oxford, 1965.

[37] Q. Wu, M. Tan, X. Li, H. Min, and N. Sun. Nmfe-sscc: Non-negative matrix factorization ensemble for semi-supervised collective classification. *Knowledge-Based Systems*, 89:160–172, 2015.

[38] P. Yang, X. Su, L. Ou-Yang, H.-N. Chua, X.-L. Li, and K. Ning. Microbial community pattern detection in human body habitats via ensemble clustering framework. *BMC systems biology*, 8(Suppl 4):S7, 2014.