Localized User-Driven Topic Discovery via Boosted Ensemble of Nonnegative Matrix Factorization^{*}

Sangho Suh $\,\cdot\,$ Sungbok Shin $\,\cdot\,$ Joonseok Lee $\,\cdot\,$ Chandan K. Reddy $\,\cdot\,$ Jaegul Choo

Received: date / Accepted: date

Abstract Nonnegative matrix factorization (NMF) has been widely used in topic modeling of a large-scale document corpus, where a set of underlying topics are extracted by a low-rank factor matrix from NMF. However, the resulting topics often convey only general, thus redundant information about the documents rather than minor, but potentially meaningful information to users. To address this problem, we present a novel ensemble method of nonnegative matrix factorization that discovers meaningful local topics. Our method leverages the idea of an ensemble model, which has shown advantages 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 extend this ensemble model further with keyword- and document- based user interaction to introduce user-driven topic discovery.

Keywords Topic modeling \cdot ensemble learning \cdot matrix factorization \cdot gradient boosting \cdot local weighting

1 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 [6]. Roughly speaking, two different topic modeling approaches exist: 1) *probabilistic models* such as probabilistic latent semantic indexing (pLSI) [20] and latent Dirichlet allocation (LDA) [6], and 2) *matrix factorization methods* such as nonnegative matrix factorization (NMF) [33].

^{*} This work is an extended version of [47]

Address(es) of author(s) should be given



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

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 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 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 domains¹ 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).

Additionally, we propose an interactive topic modeling tool that mines topics pertaining to the users' interests from the entire document corpus. For example, suppose an analyst is analyzing a large-scale dataset, such as Twitter dataset of the New York City, and is interested in understanding a particular local event, such as the New York City marathon. Although our prototypical model accomplishes

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

a thorough analysis of the dataset by providing both main and local topics of the dataset, it may not guarantee the topics that users are interested in. To supplement this limitation, we further develop a variant model that extracts topics with human intervention in the weighting process.

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 general, thus uninformative topics. In addition, we expand our work as a flexible, user-interactive method by incorporating user inputs in our boosting framework of the ensemble NMF.

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.

4. We present a topic model that extracts user-specified local topics from largescale datasets, such as Reuters news data and the information-scarce Twitter data.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our ensemble NMF approach, which can reveal diverse localized topics from text data. Section 4 shows quantitative comparison results and qualitative topic examples using various real-world datasets. Finally, Section 5 concludes the paper with future work.

2 Related Work

Since NMF was originally proposed by Paatero and Tapper [44] 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 [33]. To improve the performance and the convergence properties of NMF, many studies presented an efficient alternating nonnegative least squares (ANLS)-based framework [39,25] and its hierarchical version (HALS) [11]. In addition, Kim and Park proposed the active-set-like fast algorithms [28]. On the other hand, NMF has been applied in various manner, e.g., handling user inputs [10] and multiple data sets [23]. Many variants of NMF, such as sparse NMF [24] and orthogonal NMF [13], were also proposed using standard NMF [26].

Related to our approach, Biggs et al. [5] 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 [16] 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 [18], classification [51], and bioinformatics [52]. 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 [46]. Ensembles of maximum margin matrix factorizations (MMMF) improved the result of a single MMMF model [12]. Ensembles of the Nystrom method [32] and of the divide-and-conquer matrix factorization [40] have also been shown effective. The Netflix Prize runner-up [45] proposed a feature-weighted least squares method using a linear ensemble of learners with human-crafted dynamic weights. Lee et al. [37] proposed a stage-wise feature induction approach, automatically inducing local features instead of human-crafted features. Local low-rank matrix factorization (LLORMA) [35,36] 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.

In topic modeling, latent Dirichlet allocation (LDA) [7] is one of the most widely used methods, and researchers improved it in various ways to extract more meaningful and useful topics than LDA. Multi-grain topic modeling [48] extracts user-oriented ratable topics from user reviews. Topic modeling has also been directly integrated with sentiment analysis in order to reveal sentiments for different aspects of a product [22]. A visual analytics system TIARA [49] uses LDA-based topic analysis techniques to discover newly evolving topics. NMF has also been a popular technique in topic modeling applications. A new high-quality sentiment analysis model has been developed using nonnegative matrix tri-factorization to learn from lexical prior knowledge in sentiment classification [38].

Various interactive techniques and systems have been introduced to provide user-specified meaningful and precise topics. The work by Andrzejewski et al. [2] present interactive topic modeling to users by providing functions such as 'merging,' 'isolating,' or 'splitting' in the formation of topics. iVisClustering [34] allows one to interactively refine topic clusters generated by LDA to filter noisy data. Eddi [4] is an interactive topic browser based on clustering user's explicitly or implicitly mentioned Twitter feeds through topic analysis. ConVisIT [21] integrates LDA-based topic modeling with interactive visualization techniques in exploring long conversations from a social networking service or revising the topic model if the topic does not meet the user's needs. Bakharia et al. [3] assist qualitative content analysis of analysts by incorporating interactiveness on topic modeling algorithms.

In both LDA- and NMF-based topic modeling, most of the existing approaches extract topics from a holistic view of a document corpus. Our method, on the other

hand, extracts topics from a local point of view by considering only part of the entire corpus. In detail, our approach can be viewed as a divide-and-conquer strategy to extract local topics. Such a strategy also provides a suitable framework for userdriven topic modeling by allowing users to flexibly choose where to focus on out of the corpus. Capitalizing on this property, we propose an additional user-interactive variant to allow a user-specified keyword- and document- based topic discovery by leveraging the idea of our localized topic modeling scheme. UTOPIAN, an interactive visual analytics system suggested by Choo et al. [9], also provides user interaction with document- and keyword- induced topics. However, our topic modeling approach differs from UTOPIAN in that our model concentrates on encompassing both major and localized topics.

3 Approach

In this section, we first review standard NMF and its applications to topic modeling. Then, we formulate our method called L-EnsNMF, the gradient-boosted ensemble NMF for local topic discovery. Afterwards, we introduce iL-EnsNMF, user-driven topic discovery approach that adds keyword- and document- based user interaction to L-EnsNMF. Table 1 summarizes the notations used throughout this paper.

3.1 Preliminaries: NMF for Topic Modeling

Notation	Description				
m	Number of keywords				
n	Number of documents				
k_s	Number of topics per stage				
q	Number of stages in L-EnsNMF				
$k \ (= k_s q)$	Number of total topics				
a_r	Row vector selected from probability distribution $P_r^{(i)}(x)$				
a_c	Column vector selected from probability distribution $P_{c}^{\left(i\right)}\left(y\right)$				
$A \in \mathbb{R}^{m \times n}_+$	Input term-by-document matrix				
$P_{r}^{\left(i ight) }\left(x ight)$	Probability distribution over row indices x 's				
$P_{c}^{\left(i ight) }\left(y ight)$	Probability distribution over column indices y 's				
$\hat{W}^{(i)} \in \mathbb{R}^{m \times k}_+$	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				
\mathcal{U}_r	Set of user-selected keywords of interest				
\mathcal{U}_c	Set of user-selected documents of interest				

Table 1: Notations used in the paper



Fig. 2: Overview of the proposed ensemble approach

Given a nonnegative matrix $X \in \mathbb{R}^{m \times n}_+$, and an integer $k \ll \min(m, n)$, nonnegative matrix factorization (NMF) [33] finds a lower-rank approximation given by

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

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

$$\min_{W, H>0} \|X - WH\|_F^2.$$
⁽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 columnwise 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.

3.2 L-EnsNMF for Local Topic Modeling

3.2.1 Overview

We propose our gradient-boosted local ensemble NMF approach called L-EnsNMF. As shown in Fig. 2, our model iteratively performs three steps, (a) residual update, (b) anchor sampling, and (c) local weighting. To put it simply, residual update finds

parts that are not fully explained by NMF. Based on this finding, anchor sampling identifies particular keywords and documents that are relatively less explained. Local weighting then boosts up these unexplained parts so that they are explained in the next iterations. In the following sections, we explain our approach in more detail.

3.2.2 Ensemble NMF Approach

In our ensemble model, 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 $W^{(i)}H^{(i)}$:

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

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

Our proposed method solves this problem in a forward stage-wise manner [19], inspired by well-known ensemble learning methods in a supervised learning context such as AdaBoost [14] and gradient boosting [15]. 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

$$R^{(i)} = \begin{cases} X & \text{if } i = 1\\ \left[R^{(i-1)} - W^{(i-1)} H^{(i-1)} \right]_{+} & \text{if } i \ge 2 \end{cases}$$
(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 3.2.4.

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

$$R^{(i)} = \left[\left[\left[X - W^{(1)} H^{(1)} \right]_{+} - W^{(2)} H^{(2)} \right]_{+} \cdots - W^{(i-1)} H^{(i-1)} \right]_{+},$$
(6)

where $\hat{W}^{(i)}$ and $\hat{H}^{(i)}$ are obtained in a forward stage-wise manner, e.g., $(\hat{W}^{(1)}, \hat{H}^{(1)})$, $(\hat{W}^{(2)}, \hat{H}^{(2)})$, 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

$$\min_{W^{(i)}, H^{(i)} \ge 0, \, i=1, \dots, q} \left\| X - \sum_{i=1}^{q} W^{(i)} H^{(i)} \right\|_{F}^{2} \tag{7}$$

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

where
$$W = [W^{(1)} W^{(2)} \cdots W^{(q)}] \in \mathbb{R}^{m \times (k_s q)}_+$$
 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 [39], a coordinate [33], or a block-coordinate descent framework [26]. 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 [19]. In detail, let $f^{(i)}$ and L be

$$f^{(i)} = f\left(W^{(1)}, \cdots, W^{(i)}, H^{(1)}, \cdots, H^{(i)}\right) = \sum_{l=1}^{i} W^{(l)} H^{(l)},$$
$$L\left(X, f^{(i)}\right) = \left\|X - f^{(i)}\right\|_{F}^{2} = \left\|X - \sum_{l=1}^{i} W^{(l)} H^{(l)}\right\|_{F}^{2},$$
(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

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

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

3.2.3 Why NMF on Residual Matrices

Traditionally, a greedy approach such as the one we proposed in Section 3.2.2 can be viewed as a rank-deflation procedure for low-rank matrix factorization, which obtains low-rank factors one at a time [50]. The power method [17], which consecutively reveals the most dominant eigenvalue and vector pairs, is a representative

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

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 [17], 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 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 3.2.4, such a deflation-based method helps to reveal highly non-redundant, diverse topics

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 [31].

from the data by preventing the significant keyword shown in a particular topic from appearing in the other topics.

3.2.4 Local Weighting

In contrast to standard NMF, which discovers mostly general but less informative topics, our ensemble approach tends to identify major but general 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

$$R_L^{(i)} = D_r^{(i)} R^{(i)} D_c^{(i)}, (10)$$

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 columnwise 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

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

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

$$P_r^{(i)}(x) = \frac{\sum_{s=1}^n R^{(i)}(x,s)}{\sum_{l=1}^m \sum_{s=1}^n R^{(i)}(l,s)} \text{ for } x = 1, \cdots, m$$
(11)

$$P_{c}^{(i)}(y) = \frac{\sum_{l=1}^{m} R^{(i)}(l, y)}{\sum_{l=1}^{m} \sum_{s=1}^{n} R^{(i)}(l, s)} \text{ for } y = 1, \cdots, n.$$
(12)

In Eqs. (11) and (12), 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

$$D_r^{(i)}(l,l) = \cos(X(a_r,:), X(l,:)) \text{ for } l = 1, \cdots, m$$
(13)

$$D_{c}^{(i)}(s,s) = \cos(X(:,a_{c}), X(:,s)) \text{ for } s = 1, \cdots, n.$$
(14)

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. (10)) into Eq. (16) 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.,

$$H^{(i)} = \underset{H \ge 0}{\arg\min} \left\| W^{(i)} H - R^{(i)} \right\|_{F}^{2}.$$
 (15)

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. (11) and (12).				
$a_r \leftarrow \text{Sample a row from } P_r^{(i)}(x).$				
$a_c \leftarrow \text{Sample a column from } P_c^{(i)}(y).$				
Compute $D_{r}^{(i)}$ and $D_{c}^{(i)}$ using Eqs. (13) and (14).				
Compute $R_L^{(i)}$ using Eq. (10).				
Compute $W^{(i)}$ using Eq. (16).				
Compute $H^{(i)}$ using Eq. (15).				
end				

3.2.5 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 [30], 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

$$\left(W^{(i)}, H^{(i)}\right) = \underset{W,H \ge 0}{\operatorname{arg\,min}} \left\|R_L^{(i)} - WH\right\|_F^2.$$
 (16)

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

$$\min_{G \ge 0} \|Y - BG\|_F^2 = \sum_i \min_{\mathbf{g}_i \ge 0} \|\mathbf{y}_i - B\mathbf{g}_i\|_2^2 \tag{17}$$

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

$$\min_{\mathbf{g}>0} \|\mathbf{y} - B\mathbf{g}\|_2^2, \tag{18}$$

which is a nonnegativity-constrained least squares problem. Here, the elements of the vector **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. (18), such an optimal solution is equivalent to the solution obtained by solving an unconstrained least squares using only the passive set of variables [27], i.e.,

$$\min \|B\left(:, \mathcal{I}_p\right) \mathbf{g}_i\left(\mathcal{I}_p\right) - \mathbf{y}\|_2^2.$$
(19)

The active-set method iteratively modifies the partitioning between \mathcal{I}_a and \mathcal{I}_p and solves for Eq. (19) 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 [30] 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 [30] 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 4, 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.

3.3 iL-EnsNMF: User-Driven Topic Discovery

3.3.1 Overview

Our L-EnsNMF extracts topics by focusing on parts of the matrix that are not fully explained. While maintaining this property, we modify the above-described local weighting scheme and formulate a user-interactive variant of L-EnsNMF called iL-EnsNMF. It takes keywords as an input from the user and reveals local topics relevant to such user-selected keywords. As shown in Fig. 5, it consists of three steps, (a) residual update, (b) anchor selection, and (c) user-driven local weighting. First, the residual update finds parts that are not fully explained by NMF. Then, the weighting is decided by the user-specified set of keywords and/or set of documents. Finally, the user-driven local weighting then boosts up these user-specified parts so that they are revealed in the next iterations. In the following section, we describe iL-EnsNMF in detail.

Fig. 5: Overview of iL-EnsNMF

3.3.2 Algorithm

The main difference of iL-EnsNMF from L-EnsNMF lies in a novel scheme of userdriven selection of anchor rows or columns rather than their random sampling from Eqs. (11) and (12). That is, given the residual matrix $R^{(i)}$ at stage *i*, we apply userdriven local weighting on this residual matrix $R^{(i)}$ to obtain its locally weighted matrix $R_r^{(i)}$ or $R_c^{(i)}$, which are either row- or column-wise weighted, respectively, as

$$R_r^{(i)} = D_r R^{(i)} \text{ and}$$
(20)

$$R_c^{(i)} = R^{(i)} D_c. (21)$$

Now, we explain how to form D_r or D_c based on a user input, which is composed of particular keywords or documents that a user is interested in, as illustrated in Fig. 5. In the case a user selects keywords of interest, whose index set is represented as \mathcal{U}_r , the *l*-th diagonal element of D_r is computed as the average cosine similarity of the *l*-th row of A and those rows of A indexed by \mathcal{U}_r , i.e.,

$$D_r(l,l) = \frac{1}{|\mathcal{U}_r|} \sum_{\tilde{r} \in \mathcal{U}_r} \cos\left(A\left(\tilde{r},:\right), A\left(l,:\right)\right) \text{ for } l = 1, \cdots, m.$$
(22)

Similarly, if the user selects documents of interest, whose index set is represented as \mathcal{U}_c , the *s*-th diagonal element of D_c is computed as the average cosine similarity of the *s*-th column of *A* and those columns of *A* indexed by \mathcal{U}_c , i.e.,

$$D_c(s,s) = \frac{1}{|\mathcal{U}_c|} \sum_{\tilde{c} \in \mathcal{U}_c} \cos\left(A\left(:, \tilde{c}\right), A\left(:, s\right)\right) \text{ for } l = 1, \cdots, n.$$
(23)

Once we form the locally weighted residual matrix via the above-described weighting scheme, we iteratively perform the same process of the original L-ensNMF while fixing U_r (or U_c) until the following condition is met:

$$\frac{\left\| \left[R^{(i+d)} \left(\mathcal{U}_{r}, : \right) - W^{(i+d)} \left(\mathcal{U}_{r}, : \right) H^{(i+d)} \right]_{+} \right\|_{F}^{2}}{\left\| R^{(i)} \left(\mathcal{U}_{r}, : \right) \right\|_{F}^{2}} > \theta \text{ or }$$
(24)

$$\frac{\left\| \left[R^{(i+d)}(:,\mathcal{U}_c) - W^{(i+d)} H^{(i+d)}(:,\mathcal{U}_c) \right]_+ \right\|_F^2}{\left\| R^{(i)}(:,\mathcal{U}_c) \right\|_F} > \theta$$
(25)

where the left-hand side represents a relative residual at stage (i + d) with respect to the residual at stage i, which is the starting stage at which we chose the keywords or documents, and θ is a pre-defined parameter value. The relative residual measures how much the residual amount remains in the submatrix of $R^{(i)}$ corresponding to the user-specified keywords or documents, with respect to stage i. The less the relative residual is, the more the relevant topics would be obtained. In other words, this criterion enables the algorithm to exhaustively extract topics relevant to user-specified keywords or documents until the amount of unexplained contents about them becomes less than a particular threshold θ .

Finally, the algorithm of iL-EnsNMF is summarized in Algorithms 2 and 3.

Algorithm 2: keyword-wise iL-EnsNMF Input: Input matrix $X \in \mathbb{R}^{m \times n}_+$, $\tilde{r} \in U_r$, integers k_s and θ Output: $W^{(\tilde{i})} \in \mathbb{R}^{m \times k_s}_+$ and $H^{(\tilde{i})} \in \mathbb{R}^{k_s \times n}_+$ for $\tilde{i} = i, \cdots$ for i = 1 to m do | Compute $D_r(i, i)$ using Eq. (22). end while satisfying Eq. (24) do | Compute $R^{(i)}$ using Eq. (5). Compute R_r using Eq. (20). Compute $W^{(i)}$ using Eq. (16). Compute $H^{(i)}$ using Eq. (15). | i = i+1end

4 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.

Algorithm 3: document-wise iL-EnsNMF

4.1 Experimental Setup

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

4.1.1 Datasets

We selected the following five real-world document datasets: 1) Reuters-21578 (**Reuters**),² a collection of articles from the Reuters newswire in 1987; 2) 20 Newsgroups (**20News**),³ from Usenet newsgroups; 3) **Enron**⁴ containing 2,000 randomly sampled emails generated by the employees of Enron Corporation; 4) IEEE-Vis (**VisPub**),⁵ 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 2.

	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 2: Summary of the datasets used

 $^{^2\ {\}rm https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html}$

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

⁴ https://www.cs.cmu.edu/~./enron/

⁵ http://www.vispubdata.org/site/vispubdata/

4.1.2 Baseline Methods

We compared our method, L-EnsNMF, against various state-of-the-art methods, including standard NMF (**StdNMF**) [26],⁶ sparse NMF (**SprsNMF**) [24],⁷ orthogonal NMF (**OrthNMF**) [13],⁸ and latent Dirichlet allocation (**LDA**) [6].⁹

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.

4.1.3 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) [43], 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

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i) p(w_j)},$$
(26)

where $p(w_i, w_j)$ represents the probability of w_i and w_j co-occurring in the same document and $p(w_i)$ represents the probability of w_i occurring in our document dataset. 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.

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 $\mathcal{T} \in \{t_1, \dots, t_k\}$ and a set of documents $\mathcal{D} = \{d_1, \dots, d_n\}$, the total document coverage is defined as

$$TDC(\mathcal{T}, \mathcal{D}) = \frac{|d \in \mathcal{D} : \exists t_i \in \mathcal{T} \text{ s.t. } |w(d) \cap w_R(t_i, c_1)| \ge c_2|}{|\mathcal{D}|}, \qquad (27)$$

⁶ https://github.com/kimjingu/nonnegfac-matlab

⁷ http://www.cc.gatech.edu/~hpark/software/nmf_bpas.zip

⁸ http://davian.korea.ac.kr/myfiles/list/Codes/orthonmf.zip

⁹ http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

Fig. 6: Sensitivity analysis of topic coherence and total document coverage across various stages when 100 topics ($k_s = 2, q = 50$) are computed using VisPub dataset. Each value represents the average topic coherence and average total document coverage of k_s corresponding topics per stage. The results were obtained by computing the average values over 20 runs. The values in parentheses indicate average standard deviation of each algorithm. They represent the average of standard deviation of the corresponding values per stage.

where w(d) represents the set of words occurring in document d and $w_R(t_i, c_1)$ 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 only the better quality of topics but also the better diversity since it explains more number of documents using the same number of topics.

4.2 Quantitative Analysis

In the following, we discuss sensitivity analysis as well as quantitative comparisons of our proposed approach against other baseline methods.

4.2.1 Sensitivity Analysis

We conducted one-at-a-time sensitivity analysis using the number of stages as the varying input, as illustrated in Fig. 6. The results show that L-EnsNMF performs better than other state-of-the-art algorithms in topic coherence and total document coverage across various number of stages. Our approach, however, does not achieve the best performance in early stages-that is, prior to q = 12 and q = 15 for topic coherence and total document coverage, respectively-but improves as the stages proceed. In the case of topic coherence, as shown in Fig. 6(a), the number of stages needs to be at least q = 12 before our approach starts to perform better than other methods. Also, it is worth noting that the topic coherence values continue to increase as the number of stages grow. For total document coverage, as shown in Fig. 6(a), our approach starts to generate topics with the best total document coverage after q = 15. On the contrary to topic coherence, the total document coverage does not demonstrate increasing trend but rather consistent performance. Based on this analysis, we chose the sets of parameters, i.e., the number of stages, for the topic coherence and total document coverage experiments, which are shown at Table 3 and Table 4. Since sensitivity analysis showed no single optimal setting that works for both the topic coherence and total document coverage as well as different trends, we chose different sets of parameters for the two experiments. Among the three parameters (q = 6, 12, 24 and q = 5, 25, 50 for topic coherence)and total document coverage, respectively), the first parameter was selected at a stage where our approach does not perform the best; the second one was chosen to show where our approach starts to generate the best topic coherence and total document coverage; the last one was chosen based on where it performs the best and is double the number of stages from the previous parameter.

4.2.2 Evaluation Measures

Topic Coherence. Table 3 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 Table 3: 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 <u>underlined</u>. The standard deviation values are shown in parentheses. They represent the average of standard deviation of the corresponding values per stage.

	$k = 12 \ (k_s = 2, q = 6)$					
	Std	Sprs	Orth	LDA	L-Ens	
	NMF	NMF	NMF	LDA	NMF	
Reuters	$1.051 \ (0.343)$	$1.121 \ (0.458)$	$0.631 \ (0.771)$	1.348 (0.625)	1.315(1.144)	
20News	1.435(0.774)	1.537(0.840)	0.920(0.318)	1.685 (0.675)	2.108 (1.352)	
Enron	1.918(0.834)	1.980(0.749)	1.885(0.836)	1.778(0.558)	2.490 (1.516)	
VisPub	0.403(0.297)	0.694 (0.452)	0.389(0.295)	0.302(0.255)	1.071 (1.513)	
Twitter	1.426(0.351)	1.649 (0.706)	$1.431 \ (0.346)$	0.487(0.179)	2.761 (1.614)	
	$k = 24 \ (k_s = 2, q = 12)$					
	Std	Sprs	Orth	IDA	L-Ens	
	NMF	NMF	NMF	LDA	NMF	
Reuters	1.213(0.485)	1.408 (0.679)	0.874(0.943)	1.399(0.580)	1.640 (1.345)	
20News	1.512(0.723)	1.795(0.819)	1.000(0.342)	2.043 (0.939)	2.334 (1.403)	
Enron	1.890(0.792)	1.792(0.966)	1.886(0.790)	1.928 (0.596)	2.370 (1.387)	
VisPub	0.517(0.343)	1.040(0.661)	0.519(0.342)	0.516(0.225)	1.406 (1.644)	
Twitter	$1.654 \ (0.656)$	1.764 (0.852)	1.671(0.702)	0.442(0.367)	2.843 (1.715)	
	$k = 48 \ (k_s = 2, q = 24)$					
	Std	Sprs	Orth	IDA	L-Ens	
	NMF	NMF	NMF	LDA	NMF	
Reuters	1.349(1.349)	1.322(1.322)	1.103(1.103)	1.590 (1.590)	1.832 (1.832)	
20News	1.637(0.692)	1.864(0.730)	1.086(0.378)	2.180(0.869)	2.375 (1.486)	
Enron	1.839(0.790)	1.881(1.318)	$1.841 \ (0.788)$	$\underline{2.065}(0.637)$	2.327 (1.157)	
VisPub	0.785(0.439)	1.356 (1.348)	$0.792 \ (0.448)$	$0.734\ (0.252)$	1.882 (1.836)	
Twitter	1.591 (0.975)	1.488 (0.799)	1.731 (0.973)	0.439 (0.766)	2.958 (1.678)	

Fig. 7: Comparison of computing times for VisPub dataset. The results were obtained from the average values over 20 runs. The values in parentheses indicate average standard deviation of each algorithm. They represent the average of standard deviation of the corresponding values per stage.

Table 4: Total document coverage of VisPub based on five different methods, as defined in Eq. (27). The reported results are averaged values over 20 runs. The best performance values are shown in **bold**, and the second best ones are <u>underlined</u>. The standard deviation values (in hundredths) are shown in parentheses. They represent the average of standard deviation of the corresponding values per stage.

	$k = 10 \ (k_s = 2, q = 5)$						
c_2 in	Std	Sprs	Orth	LDA	L-Ens		
Eq. (27)	NMF	NMF	NMF		NMF		
3	0.937(0.31)	0.923(1.0)	0.940(0.6)	0.970 (0.0)	0.941(1.0)		
4	0.778(0.8)	0.746(2.4)	0.790(2.3)	0.884 (0.0)	0.821(2.9)		
5	0.496(1.7)	0.473(3.9)	0.519(3.4)	0.666 (0.0)	0.601 (4.5)		
6	0.236(1.3)	0.229(3.9)	0.256(3.3)	0.352 (0.0)	0.350(4.4)		
7	0.081(0.9)	0.083(2.7)	0.091(1.6)	0.141(0.0)	0.153 (2.9)		
8	0.021(0.2)	0.021(1.0)	0.024(0.6)	0.037(0.0)	0.047 (1.4)		
9	0.004(0.0)	0.004(0.3)	0.005(0.2)	0.005(0.0)	0.009 (0.4)		
10	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.001 (0.1)		
Avg.	0.319(0.6)	0.301(1.7)	0.328(1.4)	0.382 (0.0)	0.365(2.0)		
	$k = 50 \ (k_s = 2, q = 25)$						
	Std	Sprs	Orth	LDA	L-Ens		
	NMF	NMF	NMF	LDA	NMF		
3	0.962(0.3)	0.951(0.9)	0.963(0.4)	0.977 (0.0)	0.972(0.3)		
4	0.770(1.0)	0.717(3.6)	0.772(1.9)	0.902 (0.0)	0.892(1.7)		
5	0.428(1.4)	0.367(4.2)	0.435(2.4)	0.651(0.0)	0.689 (3.8)		
6	0.155(0.9)	0.125(2.4)	0.158(1.6)	0.336(0.0)	0.412 (3.7)		
7	$0.039\ (0.3)$	0.030(0.9)	$0.040 \ (0.5)$	0.107(0.0)	0.178 (2.3)		
8	0.007(0.1)	0.006(0.3)	0.007 (0.2)	0.028(0.0)	0.057 (1.1)		
9	0.001 (0.0)	0.001(0.0)	0.001 (0.0)	0.001(0.0)	0.012 (0.3)		
10	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.003 (0.1)		
Avg.	0.295(0.4)	0.275(1.5)	0.297(0.8)	0.375(0.2)	0.402 (1.6)		
	$k = 100 \ (k_s = 2, q = 50)$						
	Std	Sprs	Orth	LDA	L-Ens		
	NMF	NMF	NMF	LDA	NMF		
3	0.962(0.5)	0.948(0.4)	0.962(0.4)	0.979(0.0)	0.980 (0.3)		
4	0.724(1.4)	0.676(1.7)	0.722(1.3)	0.919 (0.0)	0.889(2.1)		
5	0.346(1.7)	0.303(1.4)	0.345(1.6)	0.676 (0.0)	0.669(4.6)		
6	0.111(0.9)	0.099(0.6)	0.111(0.8)	0.336(0.0)	0.397 (4.3)		
7	0.028(0.3)	0.024(0.3)	0.028(0.3)	0.105(0.0)	0.179 (2.3)		
8	0.007(0.1)	0.005(0.1)	0.007(0.1)	0.024(0.0)	0.060 (1.1)		
9	0.002(0.1)	0.001(0.0)	0.001(0.1)	0.003(0.0)	0.017 (0.4)		
10	0.000(0.0)	0.000 (0.0)	0.000(0.0)	0.000 (0.0)	0.005 (0.1)		
Avg.	0.273(0.6)	0.257(0.5)	0.272(0.1)	0.380 (0.0)	0.400 (1.7)		

other comparable methods showing equal or even better performances at times may not perform consistently in all the datasets.

In addition, Fig. 6(a) 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.

Fig. 8: Topic examples from Twitter dataset

Total Document Coverage. Table 4 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. (27) increases. Note that a large c_2 imposes a strict condition for a particular document to be explained by a topic (Section 4.1.3). 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. 7, our method runs fastest, and more importantly, it scales better than any other methods with respect to k. As discussed in Section 3.2.5, 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.

4.3 Exploratory Topic Discovery

In this section, we present diverse interesting topics uniquely found by our methods from several datasets. Fig. 8 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

(b) L-EnsNMF

Fig. 9: Discovered topics using VisPub dataset

deliver more specific and insightful information to users. For example, it discovered 'hurricane sandi'–which devastated New York City in 2012–while both 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. 9. The results from standard NMF (Fig. 9(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. [42]. 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. [8]. 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. [41].

4.4 Interactive Topic Discovery with iL-EnsNMF

In this section, we present the results of user-driven topic discovery. L-EnsNMF has strength in providing local topics that enable a more thorough, insightful

Fig. 10: Discovered topics from Reuters dataset using keyword 'korea'

Fig. 11: Discovered topics from Reuters dataset using keyword 'japan'

summary of the dataset than other topic modeling methods. The interactive topic modeling approach using iL-EnsNMF does not only inherit such traits from its original model but also allows users to find those topics of user's interest, rather than randomly chosen ones in a fully-automated manner.

Fig. 12: Relative residual values vs. stages in iL-EnsNMF

4.4.1 Keyword-Wise and Document-Wise Topic Steering

In this section, we present user-driven topic discovery examples using both keywordand document-wise iL-EnsNMF. For the former, we conducted two experiments using the keyword 'korea' and 'japan' as the user-specified keyword, respectively. For the latter, we selected the documents with the highest frequency of 'korea' and 'japan,' respectively, and assumed that these documents are those a user selected as interesting documents.

Fig. 10 shows a group of topic keywords extracted from the early and later stages using the above-described user selection of keywords and documents. Representative topic keywords include 'dollar,' 'south,' 'japan,' and 'u.s' when using keyword-wise weighting with 'korea' as a user-selected keyword, while topic keywords such as 'trade' and 'u.s' emerged as topic keywords when using document-wise weighting with the document most relevant to 'korea.' Both keyword- and document-wise weighting in the early stages showed that the prevailing issues in Korea in 1987 were mostly related to international economics. Those topics emerging from later stages, although looking less relevant at a glance, were often more focused and insightful. One interesting topic keyword that appeared at stage 11 using document-wise weighting was 'samsung' (Fig. 10(d)). Samsung, a currently Korean multinational conglomerate, was not yet a multinational company back then.

Similar experiments were done using the user-selected keyword 'japan.' Fig. 11 shows a group of keywords from the early and later stages. In the case of keywordwise weighting, keywords such as 'u.s,' 'trade,' and 'sanctions' appeared. On the other hand, when using document-wise weighting with the document most relevant to 'japan,' topic keywords such as 'trade,' 'selling,' 'tanker,' and 'sanctions' emerged. As in the previous example using the keyword 'korea,' keywords that appeared in the early stages were also related to international economics. One interesting example in this case was found using the topic keywords, 'sanction' and 'semiconductor.' The semiconductor trade conflicts between the U.S. and Japan was one of the main issues in international economics in 1987. At later stages, similar to the previous example using keyword 'korea,' topic keywords from both keyword-wise and document-wise weighting became more focused and local, but they conveyed more meanings useful to users. At stage 14 of the keyword-wise weighting, the word 'nakasone' appeared (Fig. 11(b)). Yasuhiro Nakasone was

(a) Topics extracted by iL-EnsNMF using a selected keyword 'germany'

(b) Deflate-then-Focus method with topics about 'u.s' and 'trade' removed and then those about 'germany' emphasized

Fig. 13: Deflate-then-Focus approach

the Prime Minister of Japan in 1987. At stage 13 of document-wise weighting, the word 'yeutter' appeared (Fig. 11(d)). Clayton Keith Yeutter was the United States Trade Representative in 1987.

Fig. 12 shows the relative residual measures, as defined in Eqs. (24) and (25), over stages for the above-described examples. Fig. 12(a) shows the change of relative residuals in keyword-wise weighting and Fig. 12(b) in document-wise weighting. The figure indicates that our relative residual keeps decreasing monotonically over stages.

4.4.2 Deflate-then-Focus Scenario

Suppose the user wants to extract topics using user-specified keywords but noisy keywords are often extracted because other dominant topic components prevailing in the dataset may be combined with minor topics relevant to user-specified keywords. The Deflate-then-Focus method addresses this issue by deflating the topics about the unwanted dominant keywords in advance. First, we iterate iL-EnsNMF by selecting unwanted topic keywords as the input. As the stages proceed, the dominance of the selected keyword in the dataset progressively diminishes because the interactive weighting enables the parts related to selected keywords in the residual matrix to decrease more rapidly. When a stopping criteria defined is met (Eqs. (24) and (25)), the subsequent stages begin to run using the keywords of user's interest to extract the relevant topics.

An example of this approach is illustrated in Fig. 13. In our experiment we set the stopping threshold as $\theta = 0.5$. In detail, Fig. 13(a) shows topic keywords using iL-EnsNMF with keyword-wise weighting where 'germany' was used as the user-specified keyword throughout the iteration. It can be seen that the topics about this keyword also involve other general keywords using iL-EnsNMF with keyword-wise weighting where 'germany' as the topic keywords using iL-EnsNMF with keyword-wise weighting where we selected 'u.s' and 'trade' first as unwanted user-specified keywords and 'germany' as the topic keyword of user's interest. Compared to Fig. 13(a), Fig 13(b) no longer shows unwanted keywords such as 'u.s,' 'billion,' or 'stock.' Instead, keywords that are more closely related to 'germany', such as 'linotype' and 'stolenberg,' appeared. Linotype is a German company acquired

in German Commerzbank, and Gerhard Stolenberg was the Federal Minster of Finance of Germany in 1987, respectively.

5 Conclusion

In this paper, we presented a novel ensemble approach of NMF called L-EnsNMF for high-quality local topic discovery via a gradient boosting framework and a systematic local weighting technique. L-EnsNMF 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, L-EnsNMF 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 and 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.

We also added an interaction capability to L-EnsNMF, which we call iL-EnsNMF. This method allows users to specify the interesting keywords or documents to extract their relevant topics. We demonstrated interactive topic discovery scenarios using real-world datasets, and the topics obtained through iL-EnsNMF conveys a more meaningful summary of the user-driven topics by covering both the major and local topics lying inside the dataset.

Our work is capable of summarizing the data with more contents by including both major and local topics. However, we believe that our work has certain ways to reach its full potential. In some cases, our NMF operation is vulnerable to extracting redundant topics as the stages of our computation proceeds. In iL-EnsNMF, some seemingly-irrelevant keywords or noisy data are included when extracting topics. As our future work, we plan to improve our model algorithm by tackling the constraints imposed on our model and by developing a method that can automatically detect and filter out noisy data while extracting local topics.

In addition, we also plan to expand our work to a visual topic modeling system [29] by capitalizing on the idea of our novel topic modeling approach and further expanding the interaction capabilities of our algorithm to flexibly support an extensive user-driven topic discovery.

Acknowledgments

This work was supported in part by the National Science Foundation grants IIS-1707498, IIS-1619028, 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

- N. Aletras and M. Stevenson. Evaluating topic coherence using distributional semantics. In Proc. the International Conference on Computational Semantics, pages 13–22, 2013.
- D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 25–32, New York, NY, USA, 2009. ACM.
- 3. A. Bakharia, P. Bruza, J. Watters, B. Narayan, and L. Sitbon. Interactive topic modeling for aiding qualitative content analysis. In *Proceedings of the 2016 ACM on Conference* on Human Information Interaction and Retrieval, CHIIR '16, pages 213–222, 2016.
- M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: Interactive topic-based browsing of social status streams. In *Proceedings of the 23Nd Annual ACM* Symposium on User Interface Software and Technology, UIST '10, pages 303–312, New York, NY, USA, 2010. ACM.
- 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.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research (JMLR), 3:993–1022, 2003.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993-1022, 2003.
- U. Brandes and S. R. Corman. Visual unrolling of network evolution and the analysis of dynamic discourse. *Information Visualization*, 2(1):40-50, 2003.
- 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.
 J. Choo, C. Lee, C. K. Reddy, and H. Park. Weakly supervised nonnegative matrix
- 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.
- A. Cichocki, R. Zdunek, and S.-i. Amari. Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. In *Independent Component Analysis and Signal Sep*aration, pages 169–176. 2007.
- D. DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In Proc. the International Conference on Machine Learning (ICML), pages 249–256, 2006.
- 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.
- Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.
- J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of Statistics, pages 1189–1232, 2001.
- N. Gillis and F. Glineur. Using underapproximations for sparse nonnegative matrix factorization. Pattern Recognition, 43(4):1676–1687, 2010.
- G. H. Golub and C. F. van Loan. Matrix Computations, third edition. Johns Hopkins University Press, Baltimore, 1996.
- 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.
- 19. T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.
- 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.
- E. Hoque and G. Carenini. Convisit: Interactive topic modeling for exploring asynchronous online conversations. In Proc. 20th International Conference on Intelligent User Interfaces, IUI '15, pages 169–180. ACM.
- 22. Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In Proc. the Fourth ACM International Conference on Web Search and Data Mining (WSDM), pages 815–824, 2011.

- 23. 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.
- H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating nonnegativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.
- 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.
- 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.
- 27. J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. 2008.
- 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.
- 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.
- 30. 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.
- H. W. Kuhn. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1-2):83–97, 1955.
- S. Kumar, M. Mohri, and A. Talwalkar. Ensemble nystrom method. In Advances in Neural Information Processing Systems (NIPS), pages 1060–1068, 2009.
- D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999.
- H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164, 2012.
- 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.
- 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.
- 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.
- 38. T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In Proc. the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 244–252, 2009.
- C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. Neural computation, 19(10):2756–2779, 2007.
- 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.
- M. Meyer, T. Munzner, A. DePace, and H. Pfister. Multeesum: A tool for comparative spatial and temporal gene expression data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):908–917, Nov 2010.
- S. Mukherjea, K. Hirata, and Y. Hara. Visualizing the results of multimedia web search engines. In *Information Visualization '96, Proceedings IEEE Symposium on*, pages 64–65, 122, Oct 1996.
- 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.
- 44. 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.
- J. Sill, G. Takacs, L. Mackey, and D. Lin. Feature-weighted linear stacking. Arxiv preprint 0911.0460, 2009.
- X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009:4:2–4:2, 2009.

- 47. S. Suh, J. Choo, J. Lee, and C. K. Reddy. L-ensnmf: Boosted local topic discovery via ensemble of nonnegative matrix factorization. In 2016 IEEE 16th International Conference on Data Mining (ICDM), pages 479–488, 2016.
- I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In Proceedings of the 17th International Conference on World Wide Web, WWW '08, pages 111-120, New York, NY, USA, 2008. ACM.
- 49. F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD* international conference on Knowledge discovery and data mining, KDD '10, pages 153– 162, New York, NY, USA, 2010. ACM.
- J. H. Wilkinson, J. H. Wilkinson, and J. H. Wilkinson. The algebraic eigenvalue problem, volume 87. Clarendon Press Oxford, 1965.
- 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.
- 52. 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.