Knowl Inf Syst
https://doi.org/10.1007/s10115-017-1147-9

REGULAR PAPER

Localized user-driven topic discovery via boosted ensemble of nonnegative matrix factorization

Sangho Suh[1] · Sungbok Shin[2] · Joonseok Lee[3] · Chandan K. Reddy[4] · Jaegul Choo[2]

Received: 1 March 2017 / Revised: 19 September 2017 / Accepted: 27 December 2017

Abstract Nonnegative matrix factorization (NMF) has been widely used in topic modeling of large-scale document corpora, 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 information that might be minor, but potentially meaningful to users. To address this problem, we present a novel ensemble method based on 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. The algorithm we employ to update is novel 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.

This work is an extended version of [48].

✉ Jaegul Choo
jchoo@korea.ac.kr

Sangho Suh
sh31659@gmail.com

Sungbok Shin
sb.shin.mail@gmail.com

Joonseok Lee
joonseok@google.com

Chandan K. Reddy
reddy@cs.vt.edu

1 David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada

2 Department of Computer Science and Engineering, Korea University, Seoul, South Korea

3 Google Research, Mountain View, CA, USA

4 Department of Computer Science, Virginia Tech, Arlington, VA, USA

Published online: 08 January 2018

We subsequently extend this ensemble model by adding keyword- and document-based user interaction to introduce user-driven topic discovery.

Keywords Topic modeling · Ensemble learning · Matrix factorization · Gradient boosting · 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]. Generally 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) [34].

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 such that they can explain individual documents as much as possible. Because of this characteristic, the identified topics tend to be general ones prevalent among the entire set of documents. However, these dominant topics may not provide us with 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 when 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. 1a, where standard NMF returns 'dimension' or 'reduction' as dominant keywords in most of the topics and 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 those that are minor but meaningful and important 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. 1a), our proposed method shows much more meaningful, diverse topics, thereby allowing users to develop deep insight, as seen in Fig. 1.

Additionally, we propose an interactive topic modeling tool that mines topics pertaining to the interest of users from the entire document corpus. For example, suppose an analyst is analyzing a large-scale dataset, such as Twitter dataset of New York City, and is interested in understanding a particular local event, such as New York City marathon. Although our prototypical model accomplishes a thorough analysis of the dataset by providing both main and local topics of the dataset, it may not guarantee retrieving the topics users are interested in. To supplement this limitation, we further develop a variant model that extracts topics via 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

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

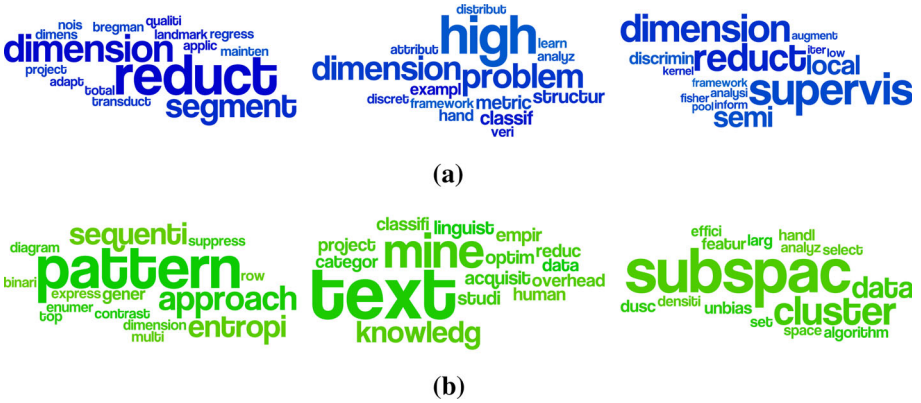


Fig. 1 Topic examples extracted from research papers in the data mining area published in 2000–2008. **a** Standard NMF. **b** L-EnsNMF

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

The remainder 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 represents the results of the quantitative comparison and qualitative topic examples using various real-world datasets. Finally, Sect. 5 concludes the paper with future work.

2 Related work

Since NMF was originally proposed by Paatero and Tapper [44] as the term positive matrix factorization, myriads of research efforts relating to NMF have been conducted. Among them, the study of Lee and Seung led to the proposal of the current popular form of NMF [34]. To improve the performance and the convergence properties of NMF, many studies presented an efficient alternating nonnegative least squares (ANLS)-based framework [25,39] and its hierarchical version (HALS) [11]. In addition, Kim and Park proposed the active-set-like fast algorithms [27]. Furthermore, NMF has been applied in various ways, e.g., handling user inputs [10] and multiple data sets [29]. Many variants of NMF, such as sparse NMF [24] and orthogonal NMF [13], were also proposed using standard NMF [28].

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 determination of an optimal submatrix for such rank-one approximation, which is computationally expensive. More recently, Gillis and Glineur [16] proposed another recursive approach known as nonneg-

ative matrix under-approximation 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 an ensemble framework in many other machine learning applications, including clustering [18], classification [53], and bioinformatics [54].

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 the previous studies were concerned with the idea of constructing an ensemble of NMF models based on a gradient boosting framework, which clearly indicated the novelty of our work.

An ensemble of general matrix factorization methods, albeit without nonnegativity constraint, has also been an active research topic in the context of collaborative filtering [47]. Ensembles of maximum margin matrix factorizations (MMMF) improved the result of a single MMMF model [12]. Ensembles of the Nystrom method [33] and of divide-and-conquer matrix factorization [40] have also been shown to be effective. The Netflix Prize runner-up [46] proposed a feature-weighted least squares method using a linear ensemble of learners with human-crafted dynamic weights. Lee et al. [36] proposed a stage-wise feature induction approach, automatically inducing local features instead of human-crafted features. Local low-rank matrix factorization (LLORMA) [37] 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) [6] 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 [49] 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 [23]. JAST [50] is a holistic fine-grained topic model that simultaneously extracts aspects and opinions by incorporating the idea of lifelong machine learning. A visual analytics system TIARA [51] 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 [35] 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] proposed ways to assist qualitative content analysis of analysts by incorporating interactiveness on topic modeling

algorithms. Recently, topic modeling has been applied to various emerging domains utilizing multimodal data analysis such as topical sentiment analysis [22], image annotation tasks [55], and analyzing the dynamics of social interactions [8,45].

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. More specifically, our approach can be viewed as a divide-and-conquer strategy to extract local topics. Such a strategy also provides a suitable framework for user-driven topic modeling by allowing users to flexibly choose the topics on which to focus in 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.[2] Afterward, we introduce iL-EnsNMF, a 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

Given a nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$ and an integer $k \ll \min(m, n)$, nonnegative matrix factorization (NMF) [34] 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

$$\min_{W,\, H \geq 0} \|X - WH\|_{\mathrm{F}}^2, \tag{2}$$

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 ith column of X, corresponds to the bag-of-words representation of document i with respect to m keywords, possibly with some preprocessing, 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 lth nonnegative column vector of W, represents the lth topic as a weighted combination of m keywords. A large value indicates a close relationship of the topic to the corresponding keyword. The ith column vector of H, $h_i \in \mathbb{R}_+^{k \times 1}$, represents document i as a weighted combination of k topics.

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

Table 1 Notations used in the paper

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 $P_c^{(i)}(y)$
$A \in \mathbb{R}_+^{m \times n}$	Input term-by-document matrix
$P_r^{(i)}(x)$	Probability distribution over row indices x's
$P_c^{(i)}(y)$	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
$R_c^{(i)} \in \mathbb{R}_+^{m \times n}$	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

3.2 L-EnsNMF for local topic modeling

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. In simple terms, 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 such that they are explained in the subsequent iterations. In the following sections, we explain our approach in more detail.

3.2.1 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 $W^{(i)}H^{(i)}$:

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

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 ith *stage* represents the ith k_s local topics discovered by the local NMF model. To achieve this approximation, we introduce an objective function in terms of the Frobenius norm as follows:

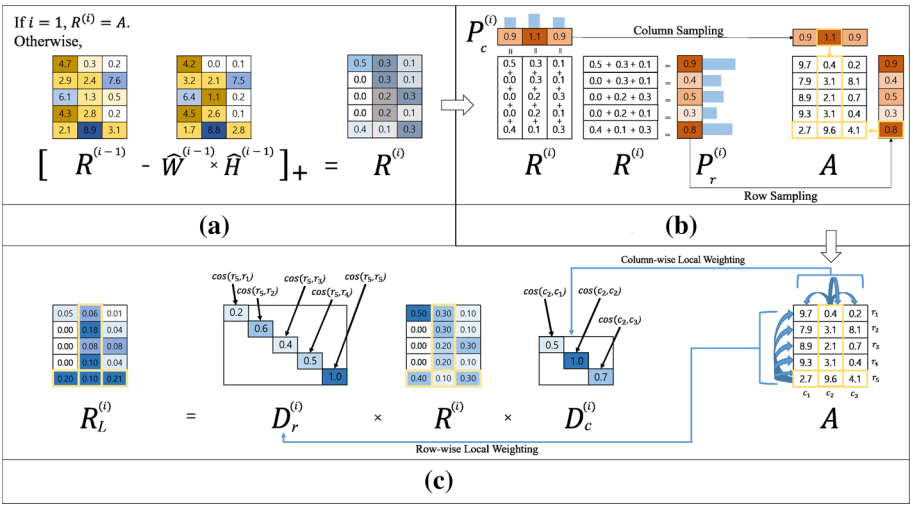


Fig. 2 Overview of the proposed ensemble approach. **a** Residual update. **b** Anchor sampling. **c** Local weighting

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

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 ith 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

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

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 $R_{\mathrm{L}}^{(i)}$ and compute $W^{(i)}$ and $H^{(i)}$ by applying NMF to $R_{\mathrm{L}}^{(i)}$ as an input matrix. More details about our local weighting scheme will be described in Sect. 3.2.3.

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

$$R^{(i)} = \Big[\Big[\Big[X - W^{(1)}H^{(1)}\Big]_+ - W^{(2)}H^{(2)}\Big]_+ \ldots - W^{(i-1)}H^{(i-1)}\Big]_+, \tag{6}$$

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

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

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

where $W = \left[\, W^{(1)}\ W^{(2)}\ \cdots\ W^{(q)} \,\right] \in \mathbb{R}_{+}^{m\times(k_{\mathrm{s}}q)}$ and $H = \begin{bmatrix} H^{(1)} \\ H^{(2)} \\ \vdots \\ H^{(q)} \end{bmatrix} \in \mathbb{R}_{+}^{(k_{\mathrm{s}}q)\times n}$.

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 [34], or a block coordinate descent framework [28]. 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

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

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

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

$$\mathbf{g}_i = \left[\frac{\partial L\left(X,\ f^{(i)}\right)}{\partial f^{(i)}} \right]_{f^{(i)}=f^{(i-1)}}$$

$$= 2\left(X - f^{(i-1)}\right) = 2\left(X - \sum_{l=1}^{i-1} W^{(l)} H^{(l)}\right).$$

Now, imposing the constraints $f^{(i)} \geq 0$ 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 $i = 1, \ldots, q$.

3.2.2 Why NMF on residual matrices

Traditionally, a greedy approach such as the one we proposed in Sect. 3.2.1 can be viewed as a rank-deflation procedure for low-rank matrix factorization, which obtains low-rank factors one at a time [52]. The power method [17], 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 [17], where the low-rank factor matrices are allowed to be both positive and negative.

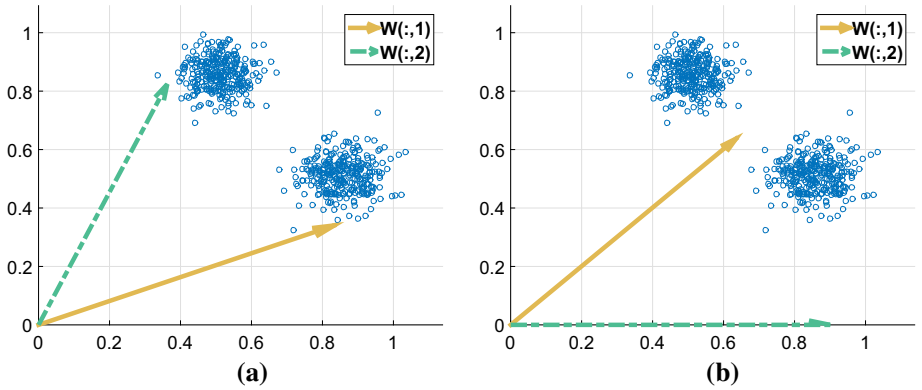


Fig. 3 Synthetic data example where $m = 2$, $k_s = 1$, and $q = 2$. **a** Standard NMF. **b** Deflation-based NMF

Generally, such a deflation method does not work for NMF, due to the limitation that the factor matrices should not contain negative elements. Figure 3 shows the comparison between standard NMF and our ensemble approach, given synthetic Gaussian mixture data in a two-dimensional feature space. As seen in Fig. 3a, 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. 3b, 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, the 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. Figure 4 shows another example of the synthetic data in which 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. 4a. In this figure, standard NMF (Fig. 4b) 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. 4c) 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, deflation-based NMF could 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 Sect. 3.2.3, 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.

3.2.3 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

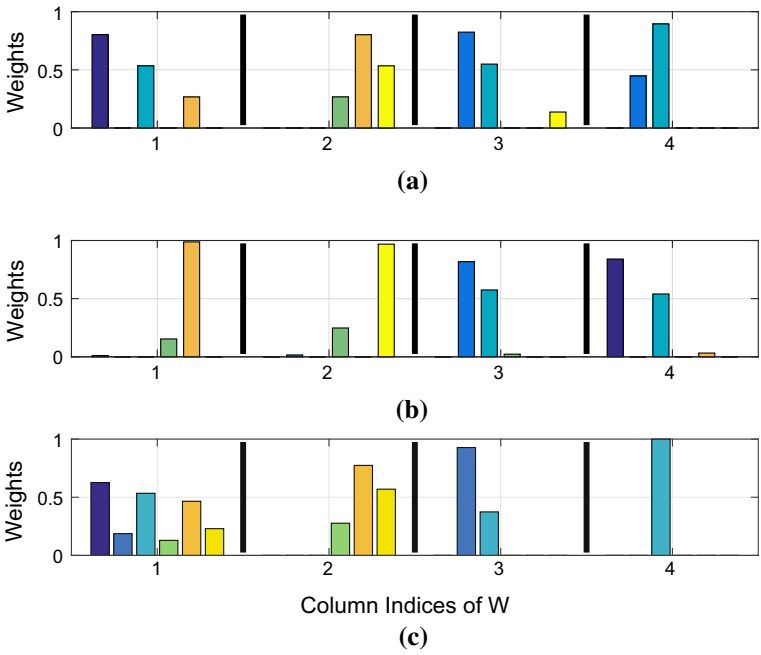


Fig. 4 Column vectors of W from synthetic data with $m = 5$, $k_s = 1$, and $q = 4$. The columns of W generated by both the standard and the ensemble NMF have been aligned with those of the ground truth W using the Hungarian method [32]. **a** Ground truth. **b** Standard NMF. **c** Deflation-based NMF

reveals interesting local topics in subsequent stages, since minor, unexplained topics can be expected to 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 $R_{\mathrm{L}}^{(i)}$ as

$$R_{\mathrm{L}}^{(i)} = D_{\mathrm{r}}^{(i)} R^{(i)} D_{\mathrm{c}}^{(i)}, \tag{10}$$

where diagonal matrices $D_{\mathrm{r}}^{(i)} \in \mathbb{R}_{+}^{m \times m}$ and $D_{\mathrm{c}}^{(i)} \in \mathbb{R}_{+}^{n \times n}$ 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

$$\begin{aligned}
&\min_{W^{(i)}, H^{(i)} \geq 0} \left\| D_{\mathrm{r}}^{(i)} \left(R^{(i)} - W^{(i)} H^{(i)} \right) D_{\mathrm{c}}^{(i)} \right\|_{\mathrm{F}}^2 \\
&= \min_{W^{(i)}, H^{(i)} \geq 0} \left\| D_{\mathrm{r}}^{(i)} R^{(i)} D_{\mathrm{c}}^{(i)} - D_{\mathrm{r}}^{(i)} W^{(i)} H^{(i)} D_{\mathrm{c}}^{(i)} \right\|_{\mathrm{F}}^2 \\
&= \min_{W_{\mathrm{L}}^{(i)}, H_{\mathrm{L}}^{(i)} \geq 0} \left\| R_{\mathrm{L}}^{(i)} - W_{\mathrm{L}}^{(i)} H_{\mathrm{L}}^{(i)} \right\|_{\mathrm{F}}^2
\end{aligned}$$

by setting $W_{\mathrm{L}}^{(i)} = D_{\mathrm{r}}^{(i)} W^{(i)}$ and $H_{\mathrm{L}}^{(i)} = H^{(i)} D_{\mathrm{c}}^{(i)}$.

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 $P_{\mathrm{r}}^{(i)}$ and $P_{\mathrm{c}}^{(i)}$ over row indices, x's, and over column indices, y's, respectively, as

$$P_{\mathrm{r}}^{(i)}(x) = \frac{\sum_{s=1}^{n} R^{(i)}(x,s)}{\sum_{l=1}^{m}\sum_{s=1}^{n} R^{(i)}(l,s)} \quad \text{for } x = 1,\ldots,m \tag{11}$$

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

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 $D_{\mathrm{r}}^{(i)}$ or $D_{\mathrm{c}}^{(i)}$, we sample from this probability distribution only a single row a_{r} and a column a_{c}, which we term 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 topics that are still unclear and reflect most of the unexplained documents.

The diagonal entries of $D_{\mathrm{r}}^{(i)}$ and $D_{\mathrm{c}}^{(i)}$ 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 lth diagonal entry of $D_{\mathrm{r}}^{(i)}(l,l)$ and the sth diagonal entry of $D_{\mathrm{c}}^{(i)}(s,s)$, respectively, as

$$D_{\mathrm{r}}^{(i)}(l,l) = \cos\left(X(a_{\mathrm{r}},:), X(l,:)\right) \quad \text{for } l = 1,\ldots,m \tag{13}$$

$$D_{\mathrm{c}}^{(i)}(s,s) = \cos\left(X(:,a_{\mathrm{c}}), X(:,s)\right) \quad \text{for } s = 1,\ldots,n. \tag{14}$$

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

Applying the localized residual matrix as described above, we plug $R_{\mathrm{L}}^{(i)}$ (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 eventually remove 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.,

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

In this manner, our approach still maintains the localized topics $W^{(i)}$ from $R_{\mathrm{L}}^{(i)}$ 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.

3.2.4 Efficient algorithm for ensemble NMF

A unique advantage of our method is that regardless of the total number of topics, k, the rank used in computing NMF at each stage, k_{s}, can be kept small while increasing the number of stages, q, i.e., $k_{\mathrm{s}} \ll (k = k_{\mathrm{s}}q)$. Hence, to efficiently solve NMF with a low value of k_{s}, we extend a recent active-set-based NMF algorithm [31], which demonstrated significantly high efficiency for a small value of k_{s}.

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

Input: Input matrix $X \in \mathbb{R}_{+}^{m \times n}$, integers k_{s} and q
Output: $W^{(i)} \in \mathbb{R}_{+}^{m \times k_{\mathrm{s}}}$ and $H^{(i)} \in \mathbb{R}_{+}^{k_{\mathrm{s}} \times n}$ for $i = 1, \ldots, q$
for $i = 1$ *to* q **do**
  Compute $R^{(i)}$ using Eq. (6).
  Compute $P_{\mathrm{r}}^{(i)}(x)$ and $P_{\mathrm{c}}^{(i)}(y)$ using Eqs. (11) and (12).
  $a_{\mathrm{r}} \leftarrow$ Sample a row from $P_{\mathrm{r}}^{(i)}(x)$.
  $a_{\mathrm{c}} \leftarrow$ Sample a column from $P_{\mathrm{c}}^{(i)}(y)$.
  Compute $D_{\mathrm{r}}^{(i)}$ and $D_{\mathrm{c}}^{(i)}$ using Eqs. (13) and (14).
  Compute $R_{\mathrm{L}}^{(i)}$ using Eq. (10).
  Compute $W^{(i)}$ using Eq. (16).
  Compute $H^{(i)}$ using Eq. (15).
end

In detail, our algorithm is built upon the two-block coordinate descent framework, which iteratively solves W while fixing H and then does this in reverse. Given a local residual matrix $R_{\mathrm{L}}^{(i)}$ 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

$$\left(W^{(i)},\ H^{(i)}\right) = \underset{W,H \geq 0}{\arg\min} \left\| R_{\mathrm{L}}^{(i)} - WH \right\|_{\mathrm{F}}^{2}. \tag{16}$$

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

$$\min_{G \geq 0} \|Y - BG\|_{\mathrm{F}}^{2} = \sum_{i} \min_{\mathbf{g}_i \geq 0} \left\| \mathbf{y}_i - B\mathbf{g}_i \right\|_{2}^{2}, \tag{17}$$

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 ith columns of G and Y, respectively. Let us consider each problem in the summation operator and rewrite it as

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

which is a nonnegativity-constrained least squares problem. Here, the elements of the vector $\mathbf{g}$ can be partitioned such that the one contains zeros and the other contains strictly positive values, and let us refer to these sets of dimension indices of the active and the passive sets as $\mathcal{I}_{\mathrm{a}}$ and $\mathcal{I}_{\mathrm{p}}$, respectively. Once we fully know $\mathcal{I}_{\mathrm{a}}$ and $\mathcal{I}_{\mathrm{p}}$ 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 [26], i.e.,

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

The active set method iteratively modifies the partitioning between $\mathcal{I}_{\mathrm{a}}$ and $\mathcal{I}_{\mathrm{p}}$ and solves for Eq. (19) until the optimal $\mathcal{I}_{\mathrm{a}}$ and $\mathcal{I}_{\mathrm{p}}$ 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 [31] 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 $2^{k_{\mathrm{s}}}$, 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 [31] proposed to build a hierarchical tree until the method obtains the number of leaf nodes as the total number of

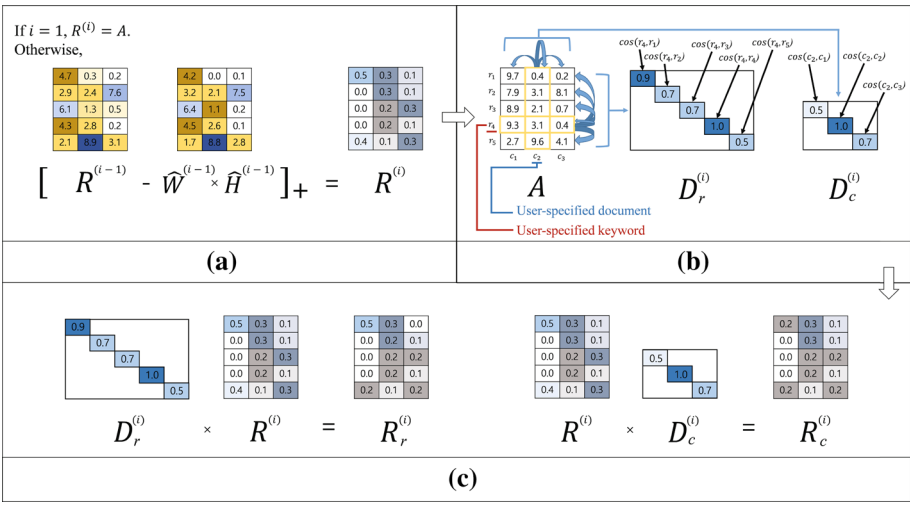


Fig. 5 Overview of iL-EnsNMF. **a** Residual update. **b** Anchor selection. **c** User-driven local weighting

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 shown in Sect. 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

Our L-EnsNMF extracts topics by focusing on those 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 named 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 such that they are revealed in the next iterations. In the following section, we describe iL-EnsNMF in detail.

3.3.1 Algorithm

The main difference of iL-EnsNMF from L-EnsNMF lies in a novel scheme of user-driven 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 user-driven local weighting on this residual matrix $R^{(i)}$ to obtain its locally weighted matrix $R_{\mathrm{r}}^{(i)}$ or $R_{\mathrm{c}}^{(i)}$, which are either row- or column-wise weighted, respectively, as

$$R_{\mathrm{r}}^{(i)} = D_{\mathrm{r}} R^{(i)} \text{ and} \tag{20}$$

$$R_{\mathrm{c}}^{(i)} = R^{(i)} D_{\mathrm{c}}. \tag{21}$$

Now, we explain how to form D_{r} or D_{c} based on user input, which is composed of particular keywords or documents that may be of interest to the user, as illustrated in Fig. 5. In the case the index set of the keywords of interest to the user is represented as $\mathcal{U}_{\mathrm{r}}$, the lth diagonal element of D_{r} is computed as the average cosine similarity of the lth row of A and those rows of A indexed by $\mathcal{U}_{\mathrm{r}}$, i.e.,

$$D_{\mathrm{r}}(l,l) = \frac{1}{|\mathcal{U}_{\mathrm{r}}|} \sum_{\tilde{r} \in \mathcal{U}_{\mathrm{r}}} \cos\left(A(\tilde{r},:), A(l,:)\right) \quad \text{for } l = 1, \ldots, m. \tag{22}$$

Similarly, if the user selects documents of interest, whose index set is represented as $\mathcal{U}_{\mathrm{c}}$, the sth diagonal element of D_{c} is computed as the average cosine similarity of the sth column of A and those columns of A indexed by $\mathcal{U}_{\mathrm{c}}$, i.e.,

$$D_{\mathrm{c}}(s,s) = \frac{1}{|\mathcal{U}_{\mathrm{c}}|} \sum_{\tilde{c} \in \mathcal{U}_{\mathrm{c}}} \cos\left(A(:,\tilde{c}), A(:,s)\right) \quad \text{for } l = 1, \ldots, n. \tag{23}$$

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 $\mathcal{U}_{\mathrm{r}}$ (or $\mathcal{U}_{\mathrm{c}}$) until the following condition is met:

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

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

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 relevant the 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 relating to 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 $X \in \mathbb{R}_+^{m \times n}$, $\tilde{r} \in \mathcal{U}_{\mathrm{r}}$, integers k_{s} and θ
Output: $W^{(\tilde{i})} \in \mathbb{R}_+^{m \times k_{\mathrm{s}}}$ and $H^{(\tilde{i})} \in \mathbb{R}_+^{k_{\mathrm{s}} \times n}$ for $\tilde{i} = i, \ldots$
for $i = 1$ *to* m **do**
  Compute $D_{\mathrm{r}}(i,i)$ 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).
  $i = i+1$
end

Algorithm 3: document-wise iL-EnsNMF

Input: Input matrix $X \in \mathbb{R}_{+}^{m \times n}$, $\tilde{c} \in \mathcal{U}_{\mathrm{c}}$, integers k_{s} and θ
Output: $W^{(\tilde{i})} \in \mathbb{R}_{+}^{m \times k_{\mathrm{s}}}$ and $H^{(\tilde{i})} \in \mathbb{R}_{+}^{k_{\mathrm{s}} \times n}$ for $\tilde{i} = i, \ldots$
for $j = 1$ *to* n **do**
  Compute $D_{\mathrm{c}}(j, j)$ using Eq. (23).
end
while *satisfying Eq.* (25) **do**
  Compute $R^{(i)}$ using Eq. (5).
  Compute R_{c} using Eq. (21).
  Compute $W^{(i)}$ using Eq. (16).
  Compute $H^{(i)}$ using Eq. (15).
  $i = i+1$
end

4 Experiments

In this section, we present extensive quantitative comparisons of our proposed approach against other state-of-the-art methods. Afterward, 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 3.10 GHz dual Intel Xeon E5-2687W processors.

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**),[3] a collection of articles from the Reuters newswire in 1987; (2) 20 Newsgroups (**20News**),[4] from Usenet newsgroups; (3) **Enron**[5] containing 2000 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 2000 randomly selected tweets generated from a specific location of New York City in June 2013. These datasets are summarized in Table 2.

[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/.

Table 2 Summary of the datasets used

	Reuters	20News	Enron	VisPub	Twitter
#Docs	7984	18,221	2000	2592	2000
#Words	12,411	36,568	19,589	7535	4212

4.1.2 Baseline methods

We compared our method, L-EnsNMF, against various state-of-the-art methods, including standard NMF (**StdNMF**) [28],[7] sparse NMF (**SprsNMF**) [24],[8] orthogonal NMF (**OrthNMF**) [13],[9] and latent Dirichlet allocation (**LDA**) [6].[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 no clear convergence criteria exist for the Gibbs sampling-based implementation of LDA, we set the number of iterations as 2000, which is one of the most common settings. Further, 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 We assess the quality of individual topics, by utilizing the point-wise mutual information (PMI) [43], which indicates the likelihood of a pair of keywords co-occur in the same document. That is, given two words w_i and w_j, PMI is defined as

$$\mathrm{PMI}\left(w_i, w_j\right) = \log \frac{p\left(w_i, w_j\right)}{p\left(w_i\right) p\left(w_j\right)}, \tag{26}$$

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 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

[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.

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

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

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 more appropriate value of this measure than the other, then one can view the set as having not only the better quality of topics, but also more extensive diversity since it explains a greater number of documents using the same number of topics.

4.2 Quantitative analysis

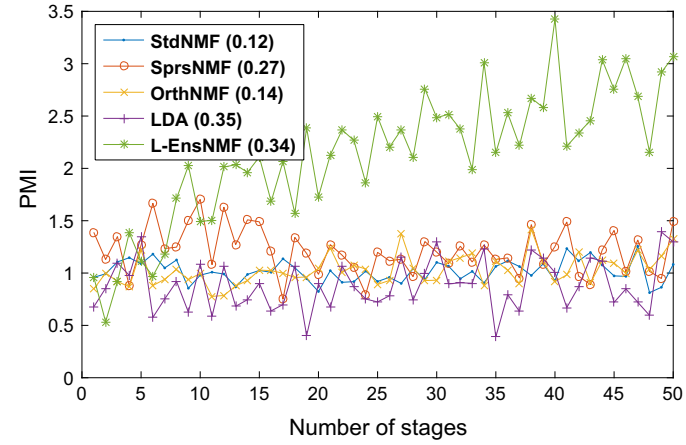
In the following, we discuss sensitivity analysis as well as quantitative comparisons of our proposed approach against other baseline methods.

4.2.1 Sensitivity analysis

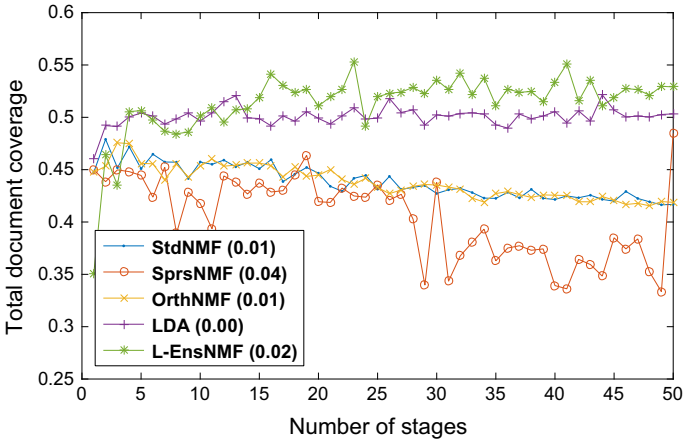
We conducted one-at-a-time sensitivity analysis using the number of stages as the varying input, as illustrated in Fig. 6. The results show that L-EnsNMF outperforms other state-of-the-art algorithms in topic coherence and total document coverage across a varying number of stages. Our approach, however, 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. 6a, the number of stages needs to be at least $q = 12$ before the performance of our approach surpasses that of other methods. Moreover, 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. 6a, our approach starts to generate topics with the best total document coverage after $q = 15$. Contrary to topic coherence, the total document coverage does not demonstrate an 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, provided in Tables 3 and 4. Since the 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 at which our approach performs less optimal; the second parameter was chosen to indicate where our approach starts to generate the best topic coherence and total document coverage; the last parameter was chosen based on where it performs the best and is double the number of stages used by 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



(a)



(b)

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 the 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. **a** Topic coherence. **b** Total document coverage

ensemble NMF is shown to maintain the highest topic coherence consistently in most of the cases. For the 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 performing equally or even more satisfactorily at times may not perform consistently on all the datasets.

In addition, Fig. 6a 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

Table 3 Comparison of topic coherence values

	Std NMF	Sprs NMF	Orth NMF	LDA	L-Ens NMF
$k = 12$ ($k_s = 2, q = 6$)					
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)
$k = 24$ ($k_s = 2, q = 12$)					
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)
$k = 48$ ($k_s = 2, q = 24$)					
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)	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)

The reported results are averaged values over 20 runs. The best performance values are shown in bold, and the second best ones are underlined. The standard deviation values are shown in parentheses. They represent the average of standard deviation of the corresponding values per stage

as the stages proceed, and eventually the quality of the topics generated by our model is much higher than with any of the 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 during the later stages in this framework are significantly more appropriate than those generated by other existing methods.

Total document coverage Table 4 shows the total document coverage results of different methods. In this table, our method is shown to be either the best or the second best method for the entire number of different 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 (Sect. 4.1.3). The fact that our method works well compared to other methods in such a strict condition signifies its important advantage of faithfully revealing 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 more efficiently than any other methods with respect to k. As discussed in Sect. 3.2.4, this computational advantage attributed to two synergetic aspects:

Table 4 Total document coverage of VisPub based on five different methods, as defined in Eq. (27)

c_2 in Eq. (27)	Std NMF	Sprs NMF	Orth NMF	LDA	L-Ens NMF
$k = 10$ ($k_s = 2, q = 5$)					
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)
$k = 50$ ($k_s = 2, q = 25$)					
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)
$k = 100$ ($k_s = 2, q = 50$)					
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)

The reported results are averaged values over 20 runs. The best performance values are shown in bold, and the second best ones are underlined. The standard deviation values (in hundredths) are shown in parentheses. They represent the average of standard deviation of the corresponding values per stage

(1) maintaining a small value for k_s regardless of the size of k is and (2) using a highly efficient NMF algorithm that performs an exhaustive search on all the possible active/passive set partitioning. These 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. Figure 8 shows the five representative topics extracted from Twitter dataset

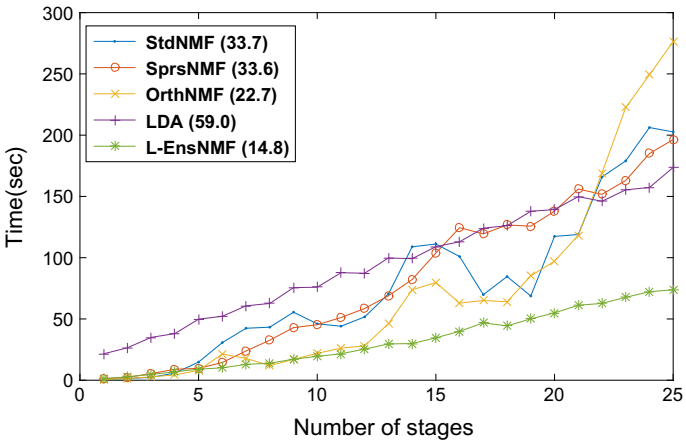


Fig. 7 Comparison of computing times for VisPub dataset. The results were obtained from the average values over 20 runs. The values in parentheses indicate average standard deviation of each algorithm. They represent the average of standard deviation of the corresponding values per stage

Fig. 8 Topic examples from Twitter dataset. **a** Standard NMF. **b** Sparse NMF. **c** Orthogonal NMF. **d** LDA. **e** L-Ens NMF

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,' and 'hahah.' In contrast, our localized ensemble NMF generates interesting keywords for its topics, e.g., 'hurricane,' 'sandy,' 'fittest,' 'survive,' and 'ireland,' which deliver more specific and insightful information to users. For example, it discovered 'hurricane sandy'—which devastated New York City in 2012—whereas neither of these words were found individually 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

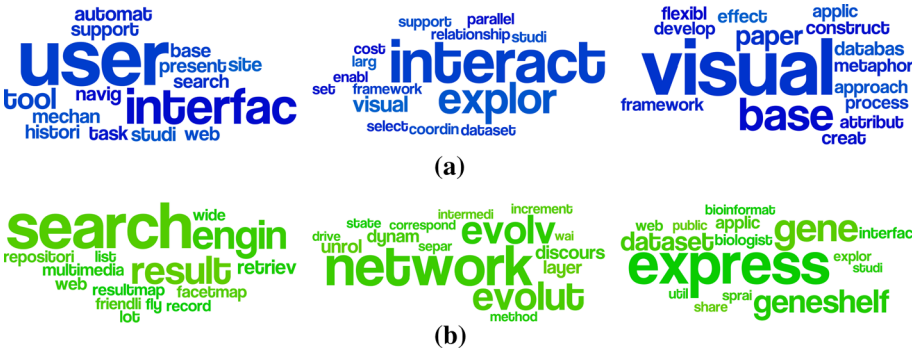


Fig. 9 Discovered topics using VisPub dataset. **a** Standard NMF. **b** L-EnsNMF

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 obtained with standard NMF (Fig. 9a) are mostly dominated by keywords that are too obvious and thus uninformative, e.g., 'visual,' 'user,' 'interface,' 'tool,' and '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. [7]. Finally, the keywords, 'gene' and 'express,' which are about biologic 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. The advantage of L-EnsNMF is its ability to provide local topics that enable a more thorough, insightful 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 topics of their interest, rather than randomly chosen ones in a fully automated manner.

4.4.1 Keyword-wise and document-wise topic steering

In this section, we present examples of user-driven topic discovery using both keyword- and 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 the user selected as interesting documents.

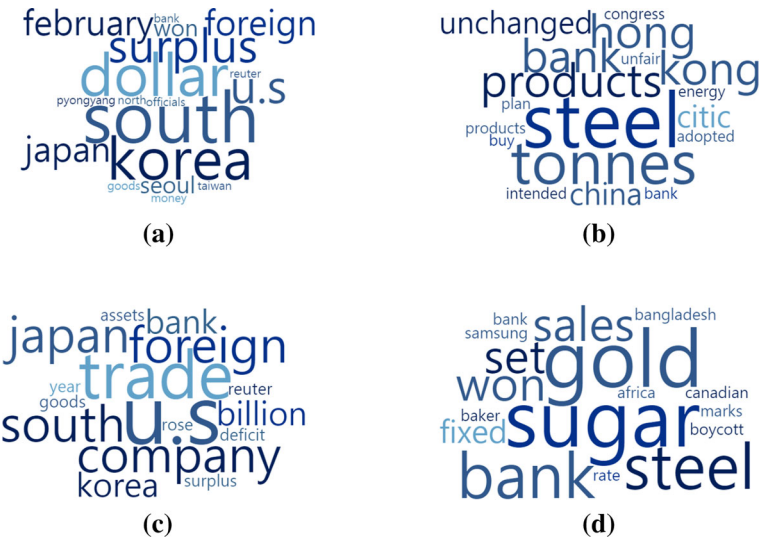


Fig. 10 Discovered topics from Reuters dataset using keyword 'korea.' **a** Keyword-wise (stages 1–2). **b** Keyword-wise (stages 11–13). **c** Document-wise (stages 1–2). **d** Document-wise (stages 11–13)

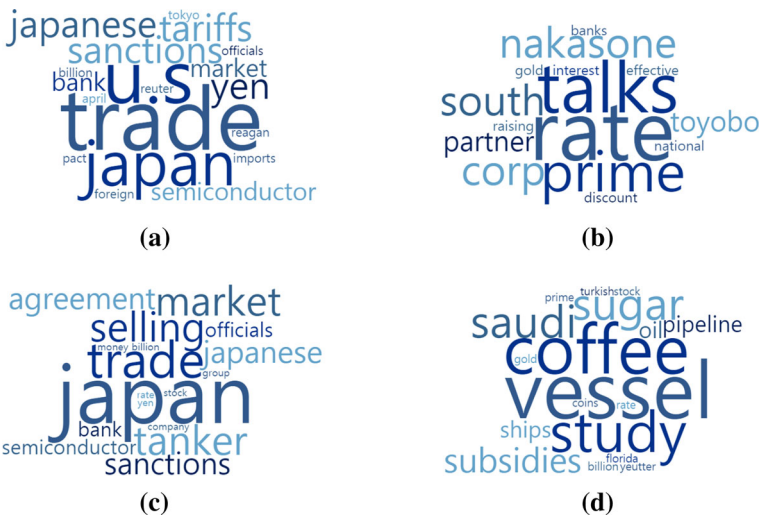


Fig. 11 Discovered topics from Reuters dataset using keyword 'japan.' **a** Keyword-wise (stages 1–2). **b** Keyword-wise (stages 13–15). **c** Document-wise (stages 1–2). **d** Document-wise (stage 13–15)

Figure 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.

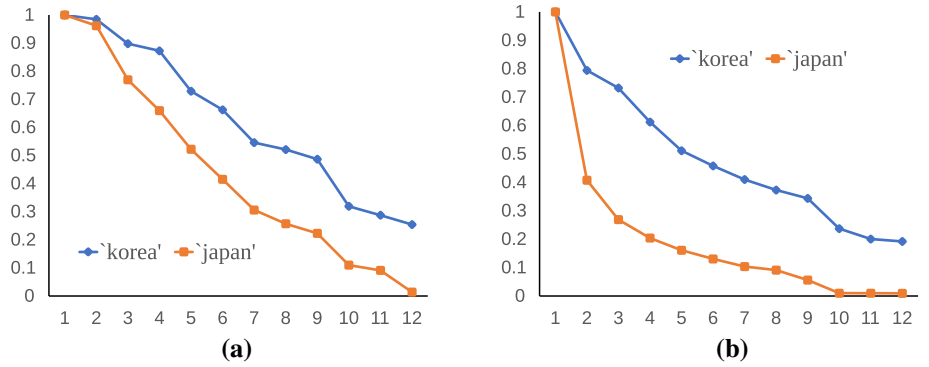


Fig. 12 Relative residual values versus stages in iL-EnsNMF. **a** Keyword-wise weighting. **b** Document-wise weighting

Those topics emerging from later stages, although appearing 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. 10d). Samsung, a currently Korean multinational conglomerate, was not yet a multinational company back then.

Similar experiments were done using the user-selected keyword 'japan.' Figure 11 shows a group of keywords from the early and later stages. In the case of keyword-wise weighting, keywords such as 'u.s,' 'trade,' and 'sanctions' appeared. However, 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 that were more useful to users. At stage 14 of the keyword-wise weighting, the word 'nakasone' appeared (Fig. 11b). Yasuhiro Nakasone was the Prime Minister of Japan in 1987. At stage 13 of document-wise weighting, the word 'yeutter' appeared (Fig. 11d). Clayton Keith Yeutter was the US Trade Representative in 1987.

Figure 12 shows the relative residual measures, as defined in Eqs. (24) and (25), over stages for the above-described examples. Figure 12a, b shows the change of relative residuals in keyword- and document-wise weighting, respectively. The monotonical decrease in relative residual values over the stages suggests that topics updated in each stage contribute to a more complete description of documents pertaining to user-specified keywords.

4.4.2 Deflate-then-focus scenario

Suppose the user intends to extract topics using user-specified keywords but noisy keywords because 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 relating to 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

Fig. 13 Deflate-then-focus approach. **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

interactive weighting enables the parts related to selected keywords in the residual matrix to decrease more rapidly. When the defined stopping criterion is met (Eqs. 24, 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 particular, Fig. 13a 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 relating to this keyword also involve other general keywords such as 'u.s' and 'trade.' Furthermore, Fig. 13b shows 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 interest to the user. Compared to Fig. 13a, Fig. 13b no longer shows unwanted keywords such as 'u.s,' 'billion,' and 'stock.' Instead, those keywords 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 Minister of Finance of Germany in 1987.

5 Conclusion

In this paper, we presented a novel ensemble NMF approach 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 performance 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 more efficiently 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 that would enable them to extract their relevant topics. We demonstrated interactive topic discovery scenarios using real-world datasets, and the topics obtained through iL-EnsNMF convey a more meaningful

summary of the user-driven topics by covering both the major and local topics contained within the dataset.

As our future work, we plan to expand our approach to an interactive topic modeling system [30] by leveraging the idea of our novel topic modeling approach and further expanding the interaction capabilities of our algorithm to flexibly support extensive user-driven topic discovery.

Acknowledgements This work was supported in part by the National Science Foundation Grants IIS-1707498, IIS-1619028, and IIS-1646881 and by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2016R1C1B2015924). 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. Aletras N, Stevenson M (2013) Evaluating topic coherence using distributional semantics. In: Proceedings of the international conference on computational semantics, pp 13–22
2. Andrzejewski D, Zhu X, Craven M (2009) Incorporating domain knowledge into topic modeling via dirichlet forest priors. In: Proceedings of the international conference on machine learning (ICML), pp 25–32
3. Bakharia A, Bruza P, Watters J, Narayan B, Sitbon L (2016) Interactive topic modeling for aiding qualitative content analysis. In: Proceedings of the ACM SIGIR on conference on human information interaction and retrieval (CHIIR), pp 213–222
4. Bernstein MS, Suh B, Hong L, Chen J, Kairam S, Chi EH (2010) Eddi: interactive topic-based browsing of social status streams. In: Proceedings of the annual ACM symposium on user interface software and technology (UIST), pp 303–312
5. Biggs M, Ghodsi A, Vavasis S (2008) Nonnegative matrix factorization via rank-one downdate. In: Proceedings of the international conference on machine learning (ICML), pp 64–71
6. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res (JMLR) 3:993–1022
7. Brandes U, Corman SR (2003) Visual unrolling of network evolution and the analysis of dynamic discourse. Inf Vis 2(1):40–50
8. Cho Y-S, Ver Steeg G, Ferrara E, Galstyan A (2016) Latent space model for multi-modal social data. In: Proceedings of the international conference on world wide web (WWW), pp 447–458
9. Choo J, Lee C, Reddy CK, Park H (2013) UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans Vis Comput Graph (TVCG) 19(12):1992–2001
10. Choo J, Lee C, Reddy CK, Park H (2015) Weakly supervised nonnegative matrix factorization for user-driven clustering. Data Min Knowl Discov (DMKD) 29(6):1598–1621
11. Cichocki A, Zdunek R, Amari S-I (2007) Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. In: Independent component analysis and signal separation, pp 169–176
12. DeCoste D (2006) Collaborative prediction using ensembles of maximum margin matrix factorizations. In: Proceedings of the international conference on machine learning (ICML), pp 249–256
13. Ding C, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix tri-factorizations for clustering. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)
14. Freund Y, Schapire R, Abe N (1999) A short introduction to boosting. J Jpn Soc Artif Intell 14(771–780):1612
15. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
16. Gillis N, Glineur F (2010) Using underapproximations for sparse nonnegative matrix factorization. Pattern Recogn 43(4):1676–1687
17. Golub GH, van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore
18. Greene D, Cagney G, Krogan N, Cunningham P (2008) Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. Bioinformatics 24(15):1722–1728
19. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin
20. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the ACM SIGIR international conference on research and development in information retrieval (SIGIR), pp 50–57

21. Hoque E, Carenini G (2015) Convisit: interactive topic modeling for exploring asynchronous online conversations. In: Proceedings of the international conference on intelligent user interfaces (IUI), pp 169–180
22. Huang F, Zhang S, Zhang J, Yu G (2017) Multimodal learning for topic sentiment analysis in microblogging. Neurocomputing 253:144–153
23. Jo Y, Oh AH (2011) Aspect and sentiment unification model for online review analysis. In: Proceedings of the ACM international conference on web search and data mining (WSDM), pp 815–824
24. Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23(12):1495–1502
25. Kim H, Park H (2008) Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. SIAM J Matrix Anal Appl 30(2):713–730
26. Kim J, Park H (2008) Sparse nonnegative matrix factorization for clustering. Georgia Institute of Technology, Georgia
27. Kim J, Park H (2011) Fast nonnegative matrix factorization: an active-set-like method and comparisons. SIAM J Sci Comput 33(6):3261–3281
28. Kim J, He Y, Park H (2014) Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J Glob Optim 58(2):285–319
29. Kim H, Choo J, Kim J, Reddy CK, Park H (2015) Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 567–576
30. Kim M, Kang K, Park D, Choo J, Elmqvist N (2017) Topiclens: efficient multi-level visual topic exploration of large-scale document collections. IEEE Trans Vis Comput Graph (TVCG) 23(1):151–160
31. Kuang D, Park H (2013) Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 739–747
32. Kuhn HW (1955) The hungarian method for the assignment problem. Naval Res Logist Q 2(1–2):83–97
33. Kumar S, Mohri M, Talwalkar A (2009) Ensemble nystrom method. In: Advances in neural information processing systems (NIPS), pp 1060–1068
34. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
35. Lee H, Kihm J, Choo J, Stasko J, Park H (2012) iVisClustering: an interactive visual document clustering via topic modeling. Comput Graph Forum 31(3 pt 3):1155–1164
36. Lee J, Sun M, Kim S, Lebanon G (2012) Automatic feature induction for stagewise collaborative filtering. In: Advances in neural information processing systems (NIPS)
37. Lee J, Kim S, Lebanon G, Singer Y, Bengio S (2016) Llorma: local low-rank matrix approximation. J Mach Learn Res (JMLR) 17(15):1–24
38. Li T, Zhang Y, Sindhwani V (2009) A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, pp 244–252
39. Lin C-J (2007) Projected gradient methods for nonnegative matrix factorization. Neural Comput 19(10):2756–2779
40. Mackey LW, Talwalkar AS, Jordan MI (2011) Divide-and-conquer matrix factorization. In: Advances in neural information processing systems (NIPS), pp 1134–1142
41. Meyer M, Munzner T, DePace A, Pfister H (2010) Multeesum: a tool for comparative spatial and temporal gene expression data. IEEE Trans Vis Comput Graph (TVCG) 16(6):908–917
42. Mukherjea S, Hirata K, Hara Y (1996) Visualizing the results of multimedia web search engines. In: Proceedings of the IEEE symposium on information visualization (InfoVis), pp 64–65, 122
43. Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic evaluation of topic coherence. In: Proceedings of the annual conference of the North American chapter of the association for computational linguistics (NAACL-HLT), pp 100–108
44. Paatero P, Tapper U (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5:111–126
45. Qian S, Zhang T, Xu C, Shao J (2016) Multi-modal event topic model for social event analysis. IEEE Trans Multimed 18:233–246
46. Sill J, Takacs G, Mackey L, Lin D (2009) Feature-weighted linear stacking. Arxiv preprint arXiv:0911.0460
47. Su X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. Adv Artif Intell 2009:4:2

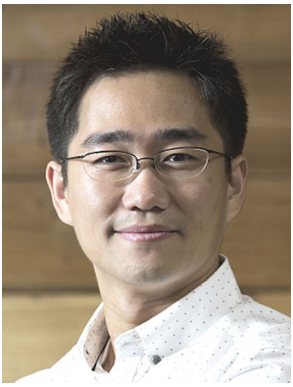
48. Suh S, Choo J, Lee J, Reddy CK (2016) L-ensnmf: boosted local topic discovery via ensemble of nonnegative matrix factorization. In: Proceedings of the IEEE international conference on data mining (ICDM), pp 479–488
49. Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. In Proceedings of the international conference on world wide web (WWW), pp 111–120
50. Wang S, Chen Z, Liu B (2016) Mining aspect-specific opinion using a holistic lifelong topic model. In: Proceedings of the international conference on world wide web (WWW), pp 167–176
51. Wei F, Liu S, Song Y, Pan S, Zhou MX, Qian W, Shi L, Tan L, Zhang Q (2010) Tiara: a visual exploratory text analytic system. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 153–162
52. Wilkinson JH, Wilkinson JH, Wilkinson JH (1965) The algebraic eigenvalue problem, vol 87. Clarendon Press, Oxford
53. Wu Q, Tan M, Li X, Min H, Sun N (2015) Nmfe-sscc: non-negative matrix factorization ensemble for semi-supervised collective classification. Knowl Based Syst 89:160–172
54. Yang P, Su X, Ou-Yang L, Chua H-N, Li X-L, Ning K (2014) Microbial community pattern detection in human body habitats via ensemble clustering framework. BMC Syst Biol 8(Suppl 4):S7
55. Zheng Y, Zhang YJ, Larochelle H (2016) A deep and autoregressive approach for topic modeling of multimodal data. IEEE Trans Pattern Anal Mach Intell (TPAMI) 38:1056–1069

Sangho Suh is a Ph.D. candidate in the David R. Cheriton School of Computer Science at the University of Waterloo. Before joining Waterloo, he earned a B.S. in Computer Science from Korea University. His research focuses on creating interactive techniques and systems that improve the way people acquire and interact with information/knowledge. His work is motivated by the desire to deepen our understanding of the nature of learning process and to propose creative solutions to make it a positive, liberating experience. Sangho received the Best Student Paper Award at IEEE ICDM conference in 2016 and a number of awards that include the University of Waterloo's International Doctoral Student Award, the GO-Bell scholarship as well as Canada's most prestigious scholarship for international students, the Ontario Trillium Scholarship.

Sungbok Shin is a M.S. student in the Department of Computer Science and Engineering at Korea University, Seoul, Republic of Korea, under the direction of Prof. Jaegul Choo. He earned his B.S. in Computer Science and Mathematics at Korea University in 2017. His research interests lie mainly in data mining and visual analytics. He is currently working on detecting anomalous events using geospatio-temporal social media data and mining user behaviors from web logs.

Joonseok Lee is a research engineer in Machine Perception at Google Research. He is mainly working on content-based video metric learning for video recommendation and annotation. He earned his Ph.D. in Computer Science from Georgia Institute of Technology in 2015, under the supervision of Dr. Guy Lebanon and Prof. Hongyuan Zha. His thesis is about local approaches for collaborative filtering, with recommendation systems as the main application. His paper 'Local Collaborative Ranking' received the best student paper award from the International World Wide Web Conference (2014). He has served as program committee in many conferences including NIPS, AAAI, WSDM, WWW, and CIKM, and journals including JMLR, ACM TIST, and IEEE TKDE. He co-organized the CVPR'17 Workshop on YouTube-8M Large-Scale Video Understanding as a program chair and served as the publicity chair for AISTATS 2015 conference. He is currently serving as a reviewer for Google Faculty Research Awards Program.

Chandan K. Reddy is an Associate Professor in the Department of Computer Science at Virginia Tech. He received his Ph.D. from Cornell University and M.S. from Michigan State University. His primary research interests are Data Mining and Machine Learning with applications to Healthcare Analytics and Social Network Analysis. He has published over 95 peer-reviewed articles in leading conferences and journals. He received several awards for his research work including the Best Application Paper Award at ACM SIGKDD conference in 2010, Best Poster Award at IEEE VAST conference in 2014, Best Student Paper Award at IEEE ICDM conference in 2016, and was a finalist of the INFORMS Franz Edelman Award Competition in 2011. He is an associate editor of the ACM Transactions on Knowledge Discovery and Data Mining and PC Co-Chair of ASONAM 2018. He is a senior member of the IEEE and life member of the ACM.



Jaegul Choo is an Assistant Professor in the Department of Computer Science and Engineering at Korea University. He has been a research scientist at Georgia Tech from 2011 to 2015, where he also received M.S. in 2009 and Ph.D. in 2013. His research focuses on visual analytics for high-dimensional data, which leverages both data mining and interactive visualization. He has been publishing in premier venues in both fields such as TVCG, VAST, CG&A, KDD, WWW, WSDM, AAAI, IJCAI, ICDM, TKDD, DMKD, ICWSM, and SDM. He earned the Best Student Paper Award at ICDM in 2016, the Outstanding Research Scientist Award at Georgia Tech in 2015 and the Best Poster Award at IEEE VAST (as part of IEEE VIS) in 2014, and he was nominated as one of the four finalists in IEEE Visualization Pioneers Group Dissertation Award in 2013.