Long Tail Constraint on Non-negative Matrix Factorization

Quanye Jia¹, Rui Liu*,¹, He Zhang¹ and Lu You¹
¹School of Computer Science and Engineering, Beihang University, Beijing, China

*corresponding author: lr@buaa.edu.cn

Abstract. The topic distribution in text generally has long tail effect, but few people do research on how to dig out long tail topics from matrix factorization. So we propose a method, this is Non-negative Matrix Factorization with Long-tail Constraint (LTNMF). LTNMF adds the soft orthogonal constraints to the feature matrix to ensure the independence of the topics on the basis of the non-negative matrix factorization. The sparse constraints and long tail constraints are added to the topic document matrix to enhance the robustness of the model and the characterization of the long tail features of the topic distribution. The combination of soft orthogonal constraints, sparse constraints and long tail constraints enables the model to extract the long tail topic information in the data and ensure the quality of the topic. We use Sougou and 20newsgroup datasets to experiment, and the results show that LTNMF can dig more topic words and improve the accuracy and the standard mutual information of clustering in text classification.

1. Introduction
In text data mining, the document is generally represented by vector, then the document set can be represented as a matrix. With the gradual increase of the amount of information, the size of the matrix has become very large and takes a long time to compute. The factorization of the matrix can not only reduce the dimension, but also compress and cluster data to help people understand the data.

The long tail distribution can be expressed as a Pareto distribution in statistics. From the distribution curve we observe that the head part of the fabric is very wide, the tail distribution is relatively flat, like a long tail, so it is called the long tail. It reflects an unbalance relationship between things, that is, a few things occupy a big head, but a large number of small things can also cause some big effects.

The topic distribution obtained by the matrix factorization algorithm is also long tail distribution. Most of the documents have the correlation with a small amount of hot topics, and other topics (long tail topics) with only a few related documents, but the long tail topics of the document has more richness than those hot topics. So it is necessary to dig out more evenly distributed topic information.

The rest of this paper is organized as follows: In Section 2, we introduce the related work on topic models and long tail theory. In Section 3, we introduce the details on LTNMF. Experiment results are presented in Section 4. At last, we give the conclusion and future work.

2. Relate work

2.1. Matrix factorization
Meth Hofmann pointed out in 1999 that topic models can be interpreted as matrix factorization [1] [2]. The commonly used matrix decomposition methods include Principal Component Analysis (PCA) [3],
Independent Component Analysis (ICA) [4], Single Value Decomposition (SVD) [5], and so on. For the principal component analysis and singular value decomposition, the matrix obtained after decomposition may contain negative numbers, which has no practical significance.

In 1999, the Non-negative Matrix Factorization (NMF) algorithm [1] was first proposed by Lee D et al. in the famous scientific journal Nature. That is to say, the matrix obtained by the original data decomposition is nonnegative. NMF has attracted many researchers’ attention, many scholars have proposed the algorithm of non-negative matrix factorization model under different constraint conditions. In 2004, Patrik proposed a non-negative matrix factorization algorithm [6] with sparse constraints. By adding L1 regularization constraints to feature topic matrix and topic document matrix, the sparse representation of the distribution of feature topics and subject documents was realized, but the classification effect could not be optimized. In 2008, Choi S [7] proposed the orthogonal non-negative matrix factorization, the full orthogonal constraint is added to feature topic matrix to keep independence between topics. He proved through the experiments that the method can achieve better results in the task of clustering and classification, and reduces the repetition ratio of the topic. Non-negative results has good explanation and the reduction effect also saves more storage space. Non-negative matrix factorization has been widely used in various fields, including text classification, image processing and clustering, and has achieved good results.

2.2. Long tail theory
In October 2004, the "connection" magazine editor Chris Anderson published "the long tail" [8] for the first time in this paper and formally put forward the theory of long tail in United States. He pointed out in the paper that under the condition of rich economy, people's attention to goods is gradually transferred from hot commodity to small public goods, which brings the market of small public goods. The aggregation of a large number of small markets has formed a large market and occupies a more important position in the whole industry. After a long time of research and practice, Anderson's view has been widely recognized, has made great success in the business, and become the focus of attention.

3. Method
3.1. Non-negative Matrix Factorization with Long-tail Constraint (LTNMF)
The matrix obtained by NMF is non-negative, NMF is used as the matrix factorization algorithm. The basic method of NMF [1] is: for a given non-negative matrix \( D \in R_{M \times N} \) and a positive integer \( k < \min\{M, N\} \), find two non-negative matrices \( U \in R_{M \times K} \) and \( V \in R_{K \times N} \) to make matrix \( D \approx UV \).

\[
\min_{U,V} f(U,V) = \frac{1}{2} \| D - UV \|_F^2 \\
\text{s. t.} \ U \geq 0, \ V \geq 0 \tag{1}
\]

Among them, \( D \) represents the feature (words) document matrix, \( U \) represents the feature topic matrix, \( V \) represents the topic document matrix, \( M \) represents the number of words in the dictionary, \( N \) represents the data set number of documents, \( K \) represents the number of topic. \( F \) represents Frobenius norm. Through statistical analysis, the weight of a small part of the topic (hot topic) in the document set is large, and most the weight of the topic (long tail topic) is small, such as the green curve shown in figure 1. In order to dig the long tail topic, we can reduce the gap between the topic weight. In the topic document matrix \( V, v_{ij} \) representing the weight of the i-th topic in the j-th document, then the r line in the matrix \( V(v_r) \) can represent the weight vector of the r-th topic in the document set, and \( v_r^2 \) (or \( v_r v_r^T \)) can be used in order to quantify this value. Due to the weight difference in the document of each subject is large, we subtract the different value of \( w_r \), to shorten the gap between the topic weight to improve the probability of mining the long tail topics. Assuming that the original topic weight is \( v_r v_r^T \), the weight of optimization is \( v_r v_r^T - w_r \) (the red curve shown in Figure 1), \( w_r \) (yellow in Figure 1) can be used by the function of characteristics of the long tail.
For a given data set, the standard non-negative matrix factorization algorithm is difficult to obtain very clear local features, often accompanied by noise, and can’t well dig the long tail topic features in data sets. According to the long tail theory, the distribution of topics in the document presents the probability distribution of Pareto distribution. Therefore, in order to better dig out the long tail information in the corpus, a reasonable long tail constraint is applied to the feature topic matrix, and the formal representation is
\[ \sum_{k=1}^{K} \| v_m^k v_n^k - w_k \|_F^2 \] (w is the weight vector of K dimension by Pareto distribution) and the soft orthogonal constraint \[ \| U^T U - I_K \|_F^2 \leq t, (t \geq 0) \] \( (I_K \) represents identity matrix of order K), the feature topic vector is approximately orthogonal. In addition, due to the large scale of training data, the feature dimension is very high, but each data sample is only related to a small amount of features. In order to represent data better, the L2 regularization constraint \( (\| V \|_F^2) \) is added to the representation vector of the data to realize sparse representation of the topic document matrix.

Thus, for the feature document matrix \( D = [d_1, d_2, ..., d_N] \), model LTNMF can be represented in the following mathematical equation (2):

\[
\min_{V, \lambda} \| D - UV \|_F^2 + \alpha \left( \| U^T U - I_m \|_F^2 \right) + \frac{\lambda}{2} \| V \|_F^2 + \frac{\beta}{4} \sum_{k=1}^{K} \| v_m^k v_n^k - w_k \|_F^2 \\
\text{s.t. } u_m^k \geq 0, m = 1, 2, ..., M; k = 1, 2, ..., K \\
\quad v_n^k \geq 0, n = 1, 2, ..., N; k = 1, 2, ..., K \\
\quad \alpha > 0, \lambda > 0, \beta > 0
\]

(2)

To determine the weight value of each topic in a document set, the Pareto distribution function with long tail features can be used, such as equation (3) (Figure 2):

\[
F(x) = Pr(X > x) = \begin{cases} 
\left( \frac{x}{x_{min}} \right)^{-k}, & x \geq x_{min} \\
1, & x \leq x_{min} 
\end{cases}
\]

(3)

Among them, x is the random variable of Pareto distribution, \( x_{min} \) is the boundary of variable x of the normal distribution and the power law distribution, also means the starting point of the power law curve. k is the shape parameter (positive number), the greater the k value, the steeper the curve descent trend.

Pareto distribution is a continuous probability distribution, and what we need is a number of discrete points in the curve to generate the weight vector of the topic, and retain the distribution characteristics of the curve. Therefore, we select an area in the power law curve (such as equation (3)), and then get the value from equal distance as the W vector, and the generating rules can be expressed as the following mathematical equation (4):
Among them, \( P_{\text{BASE}} \) is the unifying weight value of the hot topic, \( \frac{(P_{\text{MAX}} - P_{\text{MIN}}) / (K \times P_{\text{MIN}})(r - P_{\text{LONG}})}{+1} \) is the tail distribution equation of the input, which is \( x \) in the equation (3), \( P_{\text{MAX}} \) and \( P_{\text{MIN}} \) are the upper and lower bounds of \( x \) for power law curve selected in the equation (3), \( P_L \) is a shape parameter, \( K \) is the number of topic in data, \( P_{\text{LONG}} \) is the number of hot topic.

3.2. The algorithm of LTNMF

In equation (2), there are two unknown matrices of \( U \) and \( V \), which can not directly get the global optimal solution through matrix operation directly. Therefore, this paper adopts the methods commonly used by most scholars at present, to fix \( U \) for \( V \), to fix \( V \) for \( U \), so that the iteration is alternately, and then the local optimal solution of the model is obtained, and the basic procedure of LTNMF is shown in Table 1.

**Table 1.** The basic procedure of LTNMF.

| Algorithm 1: Solving algorithm of LTNMF |
|---------------------------------------|
| **Input:** \( D \in R^{M \times N}, \alpha, \beta, \lambda, P_L, P_{\text{MIN}}, P_{\text{MAX}}, P_{\text{LONG}}, P_{\text{BASE}} \) |
| **Output:** \( U \in R^{M \times K}, V \in R^{K \times N} \) |
| \( U^{(0)} \in R^{M \times K} \leftarrow \text{random}(0,1) \) |
| \( V^{(0)} \in R^{K \times N} \leftarrow \text{random}(0,1) \) |
| \( W \in R^{K \times N} \leftarrow \text{perts}(P_L, P_{\text{MIN}}, P_{\text{MAX}}, P_{\text{LONG}}, P_{\text{BASE}}) \) |
| for \( t = 1: T \) do |
| \( U = \text{UpdateU}(D, U, V) \) |
| \( V = \text{UpdateV}(D, U, V, W) \) |
| end for |
| return \( U, V \) |

We can use gradient descent method to solve matrix \( U \) and \( V \) and get the update equation as follows, \( \theta_U, \theta_V \) are gradient descent steps.

\[
U^* = U - \theta_U \frac{\partial L}{\partial U} \quad (5)
\]

\[
V^* = V - \theta_V \frac{\partial L}{\partial V} \quad (6)
\]

According to the matrix operation rule \( tr(AB) = [A]^T, tr(AB) = tr(BA), tr(A) = tr(A^T) \), the upper equation (2) can be rewritten as:

\[
L = tr((D - UV)(D - UV)^T) + \alpha \cdot tr((U^TU - I_K)(U^TU - I_K)^T) + \frac{\lambda}{2} tr(VV^T) + \beta \sum_{r=1}^{k} \|v_r v_r^T - W_r\|^2_F
\]

\[
= tr(DD^T) - 2tr(DV^TU) + tr(UVV^TU^T) + tr(U^TU^TU) - 2\alpha \cdot tr(U^TU) + \alpha \cdot tr(I_K I_K) + \frac{\lambda}{2} tr(VV^T) + \frac{\beta}{4} \sum_{r=1}^{k} \|v_r v_r^T - W_r\|^2_F \quad (7)
\]

The objective function \( L \) seeks the partial derivative of the variable matrix \( U \) and \( V \):

\[
\frac{\partial L}{\partial U} = 2(UVV^T - DV^T) + \alpha (4U^TU - 4U)
\]

\[
\frac{\partial L}{\partial V} = 2(U^TU - U^TD) + \lambda V + \beta \begin{bmatrix}
 v_1 v_1^T - w_1 & \cdots & 0 \\
 \vdots & \ddots & \vdots \\
 0 & \cdots & v_K v_K^T - W_K
\end{bmatrix}
\]

\[
(9)
\]
We can assign \( \theta_v = \frac{v}{2 U^\top UV + \lambda v + \beta \text{diag}(v_1, \ldots, v_K)} \), \( \theta_u = \frac{u}{2 U^\top UV + 4aU^\top U} \) and get the following update rules:

\[
\begin{align*}
  u_{ij} &\leftarrow u_{ij} \frac{(DV^\top + 2aU)_{ij}}{(UV^\top + 2a(U^\top U))_{ij}}, \\
v_{ij} &\leftarrow v_{ij} \frac{(U^\top D)_{ij} + \beta w_{ij} v_{ij}}{(U^\top UV)_{ij} + \lambda v_{ij} + \beta v_i v_j v_{ij}}.
\end{align*}
\]  

(10)  

(11)

4. Experiments

4.1. Experimental setup

4.1.1. Dataset. This paper selects Sogou and 20newsgroup two different news corpus for experiments.

4.1.1.1. Sogou news corpus: Sogou news corpus is a large amount of news corpus edited and annotated by Sohu. It is commonly used in Chinese classification and clustering tasks. As the size of the data set is too large, this paper selects 2500 documents randomly in 5 categories of finance, IT, health, sports and tourism, and each category has 500 documents. Then this paper pretreat the selection of data pretreatment, filter the weight of low frequency words and calculate all the TF-IDF value of the document. Then we get an original data matrix of 6413×2500 dimensional.

4.1.1.2. 20newsgroup: 20newsgroup is a commonly used data set in Natural Language Processing domain such as text classification. This article uses the version after Jason Rennie processing, which contains 18828 single annotation documents, almost evenly distributed in 20 news classes under 6 major classes. This paper uses the six major classes (alt.atheism, comp.os.ms-windows.misc, misc.forsale, rec.motorcycles, sci.electronics, talk.politics.misc) to experiment. For each class selected in the experiment, this paper selects all the documents, preprocessing and computing TF-IDF. Finally, 19183×5514 matrix is obtained.

4.1.2. Baseline methods: In order to verify the effectiveness of the proposed model LTNMF, this paper selects latent Dirichlet distribution (LDA) [9], Hierarchical Latent Dirichlet distribution (hLDA) [10], non-negative matrix factorization (NMF), orthogonal sparse non-negative matrix factorization (OSNMF) as baseline methods to verify the effectiveness of the proposed model LTNMF.

4.1.3. Parameter setup. This paper used cross validation to determine the parameters of LTNMF under better results of accuracy rate. For LTNMF model, this paper sets \( K=100, \ P_K=-3.1, \ P_{\text{MAX}}=5, \ P_{\text{MIN}}=1, \ P_{\text{LONG}}=15, \ P_{\text{BASE}}=3 \) in Sogou dataset. This paper sets \( \alpha=0.5, \ \beta=0.01, \ \lambda=0.001 \) in Sogou dataset and \( \alpha=0.05, \ \beta=0.01, \ \lambda=0.005 \) in 20newsgroup dataset. The OSNMF model is regarded as the case of LTNMF under \( K=100, \ \alpha=0.5, \ \beta=0.01, \ \lambda=0 \) condition in Sogou dataset and \( K=100, \ \alpha=0.05, \ \beta=0.01, \ \lambda=0 \) condition in 20newsgroup dataset.

For the hierarchical topic model hLDA, this paper performs about 300 tests in each experiment by cross validation, and then selects the results according to the number of topics and the accuracy of clustering, and finally determines the values of each parameter in the hLDA. In Sogou dataset, this paper sets \( \alpha=10, \ \gamma=0.1, \ \eta=10, \ \text{level}=5 \), 109 topics are obtained through 100 iterations. In 20newsgroup, this paper sets \( \alpha=5, \ \gamma=0.01, \ \eta=1, \ \text{level}=5 \), 131 topics are obtained through 100 iterations.

4.1.4. Evaluation measures:

4.1.4.1. Average number of topic words(Average): For topic models, topics can be understood as a series of related words, which can be repeated when the keywords of the two topics are the same or the
similarity is very high. Therefore, the more the number of topic words that do not repeat, the better the effect of the long tail topic mining.

4.1.4.2. Accuracy rate. According to the topic document matrix mined by the topic model algorithm, this paper uses K-means algorithm to cluster documents, and then compares the clustering categories with the actual categories, calculates the accuracy rate, and the equation is as follows:

\[
\text{Accuracy} = \sum_{c=1}^{N_c} \frac{TP_c}{TN_c}
\]  

(12)

Among them, \(N_c\) is the total number of category, \(TN_c\) is the sample number of the real category \(c\), \(TP_c\) is the sample number of the prediction category and the real category is \(c\), and the conditional \(TP_c \leq TN_c\), \(\sum_{c=1}^{N_c} TN_c = TN\) is established, so \(0 \leq \text{Accuracy} \leq 1\), and the higher the accuracy rate is better.

4.1.4.3. Standardized mutual information. Assuming that \(C\) represents a set of real clusters, \(C'\) is a set of predicted clusters, then the mutual information measure between them can be calculated by the following equation (13):

\[
MI(C, C') = \sum_{c,c' \in C, C'} p(c, c') \log \frac{p(c, c')}{p(c)p(c')}
\]

(13)

Among them, \(p(c)\) and \(p(c')\) are the probability of selecting clusters \(c\) and \(c'\) from dataset \(C\) and \(C'\), respectively. \(p(c, c')\) is the joint probability of two. Because the value of mutual information has no upper limit, it can not accurately evaluate the effect of the model. Therefore, the standardized mutual information is used to evaluate it. The equation (14) is as follows:

\[
NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}
\]

(14)

In the upper form, \(H(C)\) and \(H(C')\) represent the entropy of \(C\) and \(C'\), and it is easy to find \(0 \leq NMI(C, C') \leq 1\). \(NMI(C, C') = 0\) indicates that the two cluster sets are exactly the same. \(NMI(C, C') = 1\) indicates that the two cluster sets are independent of each other, so the greater the value of \(NMI(C, C')\), the better of the model.

4.2. Experimental results:
In this paper, we extract the feature-topic matrix \(U\) from each model in Sogou news corpus, take the first 10 and the first 20 words of every column \(u_i\) from large to small, and remove duplicates, and then calculate the average value, get the average number of keywords of each model, the results are shown in table 2.

| Model         | LDA  | NMF  | bLDA | LTNMF |
|---------------|------|------|------|-------|
| Sougou(top 10) | 5.99 | 6.92 | 7.29 | 8.23  |
| Sougou(top 20) | 10.75| 12.48| 13.11| 14.06 |
| 20newsgroup(top 10) | 5.44 | 8.36 | 8.56 | 9.19  |
| 20newsgroup(top 20) | 10.47| 16.22| 16.67| 17.16 |

The number of average topic words on the above table shows that LTNMF is the best in the 4 models and is far more than LDA. Experimental results show that LTNMF can extract more topic words compared with other methods, which shows that long tail topic mining can really improve the effect of topic mining.

After modeling by the topic model, two matrixes, the feature topic matrix \(U\) and the topic document matrix \(V\), can be obtained, which can be regarded as the projection of the original data in the low rank space after dimension reduction. This paper will apply matrix \(V\) to clustering algorithm, get
the prediction labels of each document, then compare it with the real labels, get the accuracy rate and standard mutual information results respectively as the classification effect, as shown in Figure 3 and Figure 4. It can be seen that the LTNMF with long tail constraints can improve the accuracy and the standardized mutual information of clustering. At the same time, the importance of the long tail topic information is verified.

5. Conclusion and future work
In this paper, we propose the LTNMF model, that is, added the long tail constraint to the non-negative matrix factorization. We use Sougou and 20newsgroup datasets to experiment, and the results show that LTNMF can dig more topic words and improve the accuracy and the standard mutual information of clustering in text classification.

Our proposed method is fully unsupervised. We would like to integrate supervisory information into existing long tail topic mining model LTNMF. Also, the Pareto distribution is used to fit the long tail distribution trend of the topic in the document. We would like to find another distribution function or the superposition of several distributions in place of the Pareto distribution.

References
[1] Lee D and Seung H S 2000 Algorithms for Non-negative Matrix Factorization Neural information processing systems 13 535-41
[2] Lee D and Seung H S 1999 Learning the parts of objects by non-negative matrix factorization Nature 401 788-91
[3] Abdi H and Williams L J 2010 Principal component analysis: Wiley interdisciplinary reviews computational statistics 2 433-59
[4] Comon P 1994 Independent component analysis, a new concept? Signal processing 36 287-314
[5] Golub G H and Reinsch C 1970 Singular value decomposition and least squares solutions Numerische mathematik 14 403-20
[6] Hoyer P O 2004 Non-negative matrix factorization with sparseness constraints The Journal of Machine Learning Research 5 1457-69
[7] Choi S 2008 Algorithms for orthogonal non-negative matrix factorization Proc. of the IEEE Int. Joint Conf. on Neural Networks (Piscataway: IEEE) pp 1828-1832
[8] Chris Anderson 2006 The long tail: Why the Future of Business is Selling Less of More (New York: Hyperion)
[9] Blei D M, Ng A Y and Jordan M I 2003 Latent dirichlet allocation Machine Learning Research Archive 3 993-1022
[10] Blei D M, Jordan M I, Griffiths T L and Tenenbaum J B 2003 Hierarchical topic models and the nested chinese restaurant process Int. Conf. on Neural Information Proc. Systems vol 57 (MIT Press: Cambridge MA) pp 17-24