Modeling Both Coarse-grained and Fine-grained Topics in Massive Text Data

Weifan Zhang[*], Hui Zhang[†], Yuan Zuo[†], Deqing Wang[‡]
[*]School of Computer Science, Beihang University
[†]National Engineering Research Center of S&T Resources Sharing Service, Beihang University
[‡]School of Economics and Management, Beihang University
zhangwf@buaa.edu.cn, {hzhang, zuoyuan}@nlsde.buaa.edu.cn, dqwang@buaa.edu.cn

Abstract—**Topic model has attracted much attention from investigators, as it provides users with insights into the huge volumes of documents. However, most previous related studies that based on Non-negative Matrix Factorization (NMF) neglect to figure out which topics are widespread in the documents and which are not. These widespread topics, which we refer to coarse-grained topics, have great significance for people who concentrate on common topics in a given text set. For example, after reading the massive job ads, the jobseekers are eager to learn employers' basic requirements which can be regarded as the coarse-grained topics, as well as the additional requirements that can be deemed to be the fine-grained topics. In this paper, we propose a novel method which applies two different sparseness constraints to NMF to tell coarse-grained topics and fine-grained topics apart. The experimental results of demonstrate that the new model can not only discover coarse-grained topics but also extract fine-grained topics. We evaluate the performance of the new model via text clustering and classification, and the results show the new model can learn more accurate topic representations of documents.**

Keywords—topic model; non-negative matrix factorization; text clustering; text mining

I. Introduction

With the growth of content on the internet, topics extracted from huge volumes of text documents are helpful for efficient summarizing and understanding. Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. In the past decades, great progress has been made in research area of topic models. While most of topic models focus on the human-interpretability as a whole, they pay less attention to the granularity of topics.

In some scenarios, besides the general topics, we prefer to know the coarse-grained(i.e. general) topics of corpus, for example, what are the most needed skills in the job advertisements or what is the basic knowledge of new learned area. These topics have higher document frequencies in the corpus than others, but may be less discriminating for labeled topics. This purpose of our work is to figure out how to find out both coarse-grained and fine-grained topics . We divide the whole topics into two separate parts, one is the coarse-grained topics and the other is the fine-grained topics. Coarse-grained topics appear more frequently than fine-grained topics over the corpus.

Non-negative matrix factorization (NMF)[1] is a commonly used approach in approximating high dimensional data. The non-negative property of NMF reduces the space. Sparse NMFs are also useful when we need to control the degree of sparseness in non-negative basis vectors or non-negative lower-dimensional representations. The sparseness constraint is helpful in improving the uniqueness of the decomposition along with enforcing a local-based representation. In the field of natural language processing, the text collection can be represented as a term-document matrix, whose element holds the weight (such as TF-IDF) of a term in a document. Then, the term-document matrix is approximated by the product of the two matrices which are term-topic matrix and topic-document matrix. The two lower dimension matrices characterize the relationships among terms, topics and documents. Moreover, imposing some additional constraints as regularization in NMF is a frequently used improvement to reflect the characteristics of the data more comprehensively, for example, L1 norm can approximately characterize the sparseness and L2 norm can make the data smooth.

In the paper, we propose an improved model based on NMF of term-document matrix. Our method can aggregate the general topics of the given corpus for enabling an efficient browsing and clustering. To achieve this goal, matrix factorization decomposes the term-document matrix into two parts: term-topic matrix and topic-document matrix. The sparseness of the second matrix reflect the existence among topics. With non-negative constrains on elements of matrix, NMF overcomes the shortcoming of traditional matrix decomposition (like SVD) and makes the results more interpretable. We impose L1 and L2 norm on two different parts of topic-document matrix. Specifically, we use L1 norm to characterize the fine-grained topics and L2 norm to characterize the coarse-grained topics. Due to L1 norm has the effect of making many parameters of matrix to zero which means these topics has no relation to most documents, we can found the specific topics. On the other hand, the part of matrix imposed L2 norm has more non-negative parameter, i.e., the topics are related to more documents.

Our experiments show that the proposed method can distinguish the differences of coarse-grained topics and find-grained topics, and proved that the coarse-grained topics are really widespread in the given documents. Besides, our method of non-negative matrix factorization have the feature of scaling on the distributed computing framework, for the instance, Spark (http://spark-project.org/). The approximate learning process can be decomposed into a series of mutually independent sub-optimizations. The solution procedure of optimization sub-problems can concurrently running. In this way, we can handle the big data to extract massive topics.

II. Related work

Topic model is becoming more and more important in information retrieval, machine learning and other related field, because it has been proved to be useful for relevance ranking in search, knowledge discovery and document clustering and classification. Recent studies on topic models can be categorized as probabilistic approaches and non-probabilistic approaches. Probabilistic approaches defined topics as the term probability distributions over terms and documents as data generated from mixtures of topics. PLSA[4] and LDA[5] are such models. Non-probabilistic approaches employ matrix factorization to decompose the term-document matrix into two lower dimension matrices which indicate the relevance of terms, documents and topics. LSA[6] and RLSI[7] are such non-probabilistic models.

To improve topic modeling scalability, many research have modified existing learning methods. As for LDA, AD-LDA[9], Async-CGB or Async-CVB[10] tried to implement distributed LDA algorithm. However, all the above methods need to maintain and update a dense term-topic matrix, usually in memory, which becomes a bottleneck for improving the scalability. In order to overcome the bottleneck, Wang et.al.[7] proposed RLSI whose approximate learning process can be decomposed into a series of mutually independent sub-optimizations which can be processed in parallel on large-scale data. A recent research[14] also improved the RLSI.

Regularization is a widely-used machine learning technique to prevent over-fitting. For instance, L1 regularization which uses the sum of absolute values of parameters, can make the matrix more sparse and cause many parameters to be zero[22]. While L2 regularization that uses the sum of squares of parameters has the effect of smooth the data to effectively deal with over-fitting. As a result, we found L2 regularization is fit for describing the coarse-grained topics and L1 regularization is suitable for characterizing the fine-grained topics.

Most of existing approaches rely on Singular Value Decomposition (SVD), and consequently have the following limitations: these works need to either assume that each document contains only one topic, or else can only recover the span of the topic vectors instead of the topic vectors themselves. However, NMF is a matrix factorization technique that factorizes the non-negative data matrix into two non-negative matrices. NMF has shown good performance and much work has been done in both applying NMF to different problem areas as well as on studying NMF itself[2]. Aside from the original use of NMF for learning parts of images[1], NMF has shown superior performance in document clustering[12]. The theoretical study[13] have shown the equivalence between NMF with other clustering algorithms, including K-means, Spectral Clustering and PLSA, with additional constraints.

As far as we know, the study of modeling the breadth of topics are rarely based on NMF methods. Chemudugunta, et al. [3] proposed a probabilistic model that accounts for both general and specific aspects of documents. It can be considered as an extension of LDA model and proved to have improvement on querying the specific words. In [17,18], the studies motivate by comparing the text of two or more category and mining the common or specific aspects of the text. However, our model focus on the study of dividing the whole topics into two parts, i.e., coarse-grained topics and fine-grained topics in the way of NMF. The purpose of our work is to help people effectively comprehend the massive documents.

III. Method

A. Problem Formulation

Given a text corpus with N documents, in which each document was segmented, filtered stopwords and stemmed. After that, a document can be represented as an M-dimensional vector *d*, where the *m^th^* element indicates the TF-IDF score of the *m^th^* term. The N documents are then represented in an $M \times N$ term-document matrix *D*, in which each row corresponds to a term and each column corresponds to a document. Topics are defined over terms in the document collection and can be represented as an M-dimensional vector *u*, where the m^{th} elements denotes the weight of the m^{th} term in the topic.

TABLE I. The Approximate Learning Process

```
Algorithm 1:
Input: $D \in R^{M\times N}, K, k_f, \lambda_1, \lambda_2, T$
Output: $U \in R^{M\times K}, V \in R^{K\times N}$
$V^{(0)} \in R^{K\times N} \leftarrow random(0,1)$
for t=1:$T$ do
    $U^{(t)} \leftarrow UpdateU\left(D, V^{(t-1)}\right)$
    $V^{(t)} \leftarrow UpdateV\left(D, U^{(t)}\right)$
end for
return $U^{(t)}, V^{(t)}$
```

B. Our model

Document *D* is approximated by the product of *U* and *V*, so that the term-document matrix *D* can be projected into a *K*-dimensional topic space with the least information loss in which each axis corresponds to a topic. *K* topics can be summarized into an $M \times K$ term-topic matrix $U = [u_1, u_2, \cdots, u_k]$, and the topic-document matrix is $V = [v_1, v_2, \cdots, v_N]$, and each document is represented as a linear combination of the topics. We divide the *V* matrix into two parts: one is $V^{(\alpha)} = [v_1, v_2, \cdots, v_{k_f}]^T$ and the other is $V^{(\beta)} = [v_{k_f+1}, v_{k_f+2}, \cdots, v_K]^T$ (k_f is the number of coarse-grained topics). We suggest L2 norm on $V^{(\alpha)}$ and L1 norm on $V^{(\beta)}$. In the way, we can obtain coarse-grained topic-document matrix $V^{(\alpha)}$ and fine-grained topic-document matrix $V^{(\beta)}$. Formally, given a document collection $D = [d_1, d_2, \cdots, d_N]$ as the data matrix of non-negative elements, our model amounts to solving the following optimization problem:

$$\min_{U,V} \| D - UV\|_{\hat{o}}^2 + \lambda_1 \| V^{(\alpha)}\|_2^2 + \lambda_2 \| V^{(\beta)}\|_1^1 \qquad (1)$$

s.t. $\lambda_1 > 0, \lambda_2 > 0, U \geq 0, V \geq 0, k_f$ can be fixed accordingly

$U \in R^{M\times K}$ is the term-topic matrix and $V \in R^{K\times N}$ is the topic-document matrix. $\|\cdot\|_F$ is the Frobenius norm. Specifically, each element u_{ij} of U indicates the degree of association of term i with the topic j, and v_{ij} of V denotes the relation between topics and documents. K is a pre-specified parameter denoting the dimension of reduced space and also denotes the number of desired topics, and k_f is the number of coarse-grained topics. We optimize the function in (1) by alternately minimizing it with respect to term-topic matrix U and topic-document matrix V. This procedure is summarized in TABLE I.

TABLE II. UPDATE OF MATRIX U

```
Algorithm 2: UpdateU
Input: $D \in R^{M\times N}, V \in R^{K\times N}$
Output: $U \in R^{M\times K}$
$R = VD^T$
$S = VV^T$
for i = 1:$M$ do
    $\overline{u}_i = random(0,1)$
    repeat
        for j = 1:$K$ do
            $a \leftarrow S_{jj}$
            $b \leftarrow R_{ji} - \sum_{r=1,r\neq j}^{K} S_{jr}\overline{u}_{ir}$
            $\overline{u}_{ij} = \max\{0, \frac{b}{a}\}$
        end for
    until convergence
end for
return $U$
```

1) Update of Matrix U:

Holding matrix $V = [v_1, v_2, \cdots, v_N]$ fixed, the update of U amounts to solving the following optimization problem:

$$\min_U \| D - UV \|_6^2 \quad s.t. \quad u_{ij} \geq 0$$

The above problem can be solved as follow:

$$\min_{\overline{u}_i} \sum_{i=1}^{M} \| \overline{d}_i - \overline{u}_i v \|_2^2$$

which consists of M optimization sub-problems. The solution procedure of sub-problems can be processed in parallel. $\overline{d}_i$ is the i^{th} row vector of matrix D. $\overline{u}_{i/j}$ is the row vector whose elements are the i^{th} row of matrix U without the j^{th} element and $\overline{v}_{/j}$ is the matrix with j^{th} column removed.

$$\min_{\overline{u}_i} \| \overline{d}_i - \overline{u}_{i/j}\overline{v}_{/j} - u_{ij}\overline{v}_j \|_2^2, \quad s.t. \quad u_{ij} \geq 0, i = 1, 2, \cdots, M$$

We can rewrite the object function in above equation.

$$L(\overline{u}_i) = \| \overline{d}_i - \overline{u}_{i/j}\overline{v}_{/j} - u_{ij}\overline{v}_j \|_2^2$$

$$= (\overline{v}_j \overline{v}_j^T) u_{ij}^2 - 2u_{ij}\overline{v}_j(\overline{d}_i - \overline{u}_{i/j}\overline{v}_{/j})^T + const$$

Let $S = VV^T$, $R = VD^T$, the equation can be solved as follows:

$$a = \overline{v}_j \overline{v}_j^T = (VV^T)_{jj} = S_{jj} > 0$$

$$b = u_{ij}\overline{v}_j(\overline{d}_i - \overline{u}_{i/j}\overline{v}_{/j})^T$$

$$= (VD^T)_{ji} - \sum_{r=1,r\neq j}^{K} (VV^T)_{jr}\overline{u}_{ir} = R_{ji} - \sum_{r=1,r\neq j}^{K} S_{jr}\overline{u}_{ir}$$

$$u_{ij} = \max\{0, \frac{b}{a}\}$$

The whole algorithm of update U was summarized in TABLE II.

TABLE III. UPDATE OF MATRIX V

```
Algorithm 3: UpdateV
Input: $D \in R^{M\times N}, U \in R^{M\times K}$
Output: $V \in R^{K\times N}$
$R = U^T D$
$S = U^T U$
for i=1:$N$ do
    $v_i = random(0,1)$
    repeat
        for j = 1:$K$ do
            $a \leftarrow S_{jj}$
            $b \leftarrow R_{ji} - \sum_{r=1,r\neq j}^{K} S_{jr}\overline{u}_{ir}$
            if $(1 \leq j \leq k_f)$
                $v_{ij} = \max\{0, \frac{b}{\lambda_1 + a}\}$
            else
                $v_{ij} = \max\{0, \frac{b - \frac{\lambda_2}{2}}{a}\}$
            endif
        end for
    until convergence
end for
return $V$
```

2) Update of Matrix V:

Holding matrix $U = [u_1, u_2, \cdots, u_k]$ fixed, the Update of V can also be decomposed into N optimization problems, with each corresponding to one v_n and can be solved in parallel.

$$\min_{v_i} \sum_{i=1}^{N} \| d_i - uv_i \|_2^2 + \lambda_1 \sum_{i=1}^{N} \| v_i^{(\alpha)} \|_2^2 + \lambda_2 \sum_{i=1}^{N} \| v_i^{(\beta)} \|_1^1$$

$$s.t. \quad \begin{cases} \lambda_1 > 0, \lambda_2 > 0, \\ 0 < k_f \leq K \\ v \geq 0 \quad (v^\alpha \geq 0, v^\beta \geq 0) \end{cases}$$

The procedure is similar as updating matrix *U*.

$$\min_{\{v_i\}} \| d_i - Uv_i \|_2^2 + \lambda_1 \| v_i^{(\alpha)} \|_2^2 + \lambda_2 \| v_i^{(\beta)} \|_1^1$$

$$= (u_j^T u_j) v_{ij}^2 - 2 v_{ij} u_j^T (d_i - u_{/j} v_{i/j})^T + h\lambda_1 v_{ij}^2 + (h-1)\lambda_2 v_{ij} + const$$

$$h = \begin{cases} 0, & if\, 1 \leq j \leq k_f \\ 1, & if\, k_f < j \leq K \end{cases}$$

K is the number of the topics, which is compound of k_f coarse-grained topic and (K-k_f) fine-grained topics. With the same substitute as the procedure of update *U*, let $S = VV^T$, $R = VD^T$, $a = S_{jj}$, $b = R_{ji} - \sum_{r=1, r\neq j}^{K} S_{jr} \overline{u}_{ir}$, the v_{ij} can be calculated as follows:

$$v_{ij} = \begin{cases} \max\{0, \dfrac{b}{\lambda_1 + a}\}, & if\, 1 \leq j \leq k_f \\ \max\{0, \dfrac{b - \frac{\lambda_2}{2}}{a}\}, & if\, k_f < j \leq K \end{cases}$$

The algorithm for updating matrix *V* is summarized in TABLE III.

IV. EXPERIMENT

In the section, we introduce the datasets for evaluation and reveal the results of experiment on the different metrics. We firstly investigate the quality of extracted topics and then measure the sparseness of the topic-document matrix *V*. At last, we compare the performance of VSM, LDA and our model with various parameters in terms of clustering and classification.

A. Datasets

1) 20Newsgroups:

The 20 Newsgroups data set is a popular dataset for experiments in text mining area, such as text classification and text clustering. The dataset is a collection of newsgroup documents, partitioned evenly across 20 different groups. According to the similarity between the twenty sub-classes, the 20 newsgroups are also categorized into six super-classes, i.e., science, computers, recreation, religion, politics and forsale.

The dataset was removed stopwords and three parts which contain little semantic information, i.e., headers, footers and quotes, and then kept long documents which contained 30 words or more. After that, there are 9,711 documents left to calculate TF-IDF score matrix. The term-document matrix holds 10,022 unique words and 659,394 non-zero elements.

2) Wikipedia

As known to all, Wikipedia is an abundant corpus. Xuan-hieu Phan et al.[15] crawled more than 70,000 Wikipedia documents which covered a lot of concepts and domains. To ease the computational of comparison, we choose the first 10,000 documents for our experiment.

3) International Organization Recruitment

We crawled the public recruitment information of some International Organizations, which are WHO (World Health Organization), UNESCAP (United Nations Economic and Social Commission for Asia and the Pacific) and other a dozen of organizations. These job ads are available on their websites, and then put together on our site which have only shown the job list by far. Furthermore, we desired to mine some more useful information for job seeker.

TABLE IV. TOPICS OF COMPUTER SUPERCLASS OF 20NEWSGROUP

Topics	Words
1	good pretty things luck worth fast fairly
3	motif toolkit openlook api free linux unix
5	price buy order company sell purchase sales
7	draw mydisplay drawing line bitmap xdrawline background
9	problem fine fix works solution work occur
20	mac iisi advance internal iici external platform
30	floppy boot diskette drive floptical light cmos
40	program call write small crash execute running
60	email address reply send post information info
80	int null char return void printf static
100	mouse microsoft button move ball motion logitech

TABLE V. TOPICS OF INTERNATIONAL ORGANIZATION RECRUITMENT

Topics	words
1	legal justice advisor law chief rule technical legislation right agreement commercial clinic
3	secretariat officer meeting documentation electronic communication environmental
5	uncdf[1] mobile job banking microfinance roadmap investment capital finance myanmar
7	woman gender empowerment equality right girl violence peace cultural humanitarian
9	client chemical identify programme industry issue review mercury metal technology opening
20	ebola outbreak response emergency surge logistics ipc[2] vacancy guinea country liberia
30	electronics beam accelerator control engineer digital design circuit instrumentation
40	financial finance accounting budget award payment expenditure transactions voucher
60	medicine health medical rht pharmaceutical emp vaccine clinical drug quality diagnostics
80	data statistics analysis statistical information survey indicator epidemiology biostatistics
100	security undss[3] lsa[4] residential syrian incident somalia safety risk training military

[1] UNCDF: United Nations Capital Development Fund

[2] IPC: Infection prevention and control

[3] UNDSS: United Nations Department of Safety and Security

[4] LSA: Local Security Assistant

B. Topic Evalution

We first investigate the readability of topics that we randomly select from the whole 150 topics, and the topic index as the first column indicates in TABLE IV. Each topic is displayed by the top seven terms. The first 10 topics we have found can be considered as coarse-grained topics of the corpus, as a result of choosing the super class of computer, these topics are also regarded as common topics of the computer field. For example, the No.1 topic's terms are used for describing the quality of software or other things, which should be more general than other words. The No.5 topic is about transaction and the No.7 topic is about graphics. The No.9 talks about problems of computer program. Obviously, all these ten topics can be proved to be the coarse-grained topic. The subsequent ninety topics, by contract, talk about more specific things, like mac, disk and email. Notably, the eightieth topic consists of all program language words which is beyond all doubt a fine-grained topic.

Besides the 20newsgroup dataset, we apply our method on another dataset. We crawled a lot of job ads on the internet to manifest the motivation that we talk about in the abstract. In TABLE V. the first ten topics is about more common things which reflect various aspect of job requirements, such as legal advisor, secretary, feminism. All these common topics give us a glimpse of the massive job ads and save our time to review the requirements. It is really helpful for people who are seeking for an international group job. Meanwhile, the other topics also demonstrate that our model is an effective method to model topics. Due to the limit of space, the similar results of Wikipedia are not listed here.

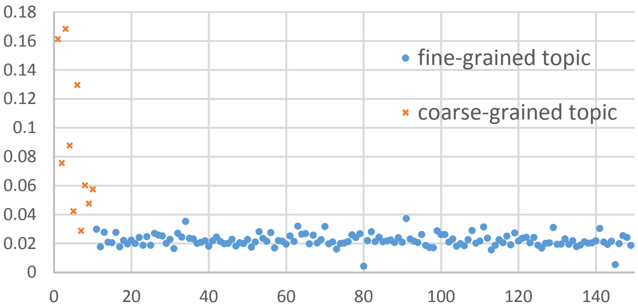

Fig. 1. Terms' document frequency of various topics

C. Topic Sparseness

We calculate the terms' document frequency (*TDF*) of the topics as follows:

$$TDF_{(\mathrm{i})} = \frac{1}{n}\sum_{k=1}^{K}\frac{1}{N}t_{df}^{(\mathrm{k})}, V_{ik} \geq \theta$$

The above equation shows the i-th topic's *TDF* value is related to these parameters, where *N* is the whole documents number, *K* is the number of the term *t* whose weight in *V* matrix is above the threshold *θ*. As Fig. 1 illustrate that *TDF* value of the first ten topics are much larger than the subsequent topics and that means the coarse-grained topics exists much more widely in the corpus than fine-grained topics.

We compare the two component's sparseness of *V* matrix affected by λ_1 and λ_2. The calculation of sparseness just count the elements of matrix which larger than 10^{-4}. Fig. 2 shows that the parameter of non-negativity constrains on the *V* matrix worked. With the increasing of λ_1, *V* matrix has nearly same sparseness without consideration of random error. However, the variable value of λ_2 has much influence on the lower part of matrix. As we stated in the Fig. 3, the larger λ_2 can make the lower part of *V* matrix more sparseness. In the way, we figure out the low frequency appearance topics.

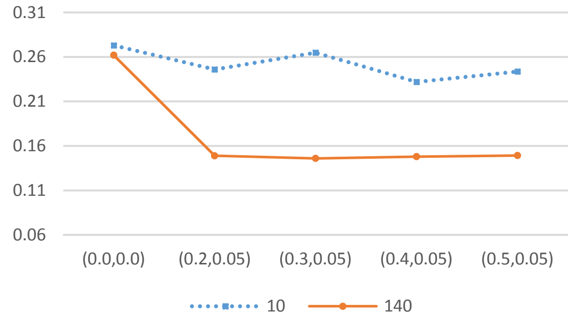

Fig. 2. Spareness of wikipedia corpus with different parameter(λ_1 , λ_2)

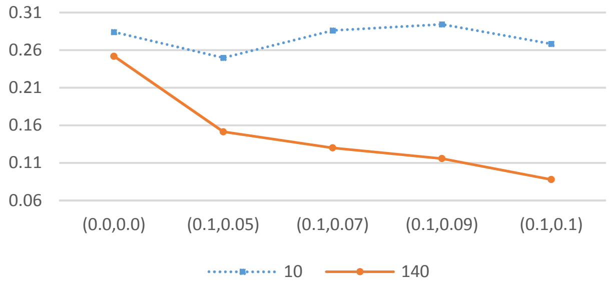

Fig. 3. Spareness of 20newsgroup corpus with different parameter(λ_1 , λ_2)

D. Clustering and Classification

We utilize 20 newsgroup corpus to evaluate the performance of clustering which uses features of our model, LDA and VSM. In order to making the result more clearly, we take one of super class of this corpus, computer. Our model preform much better than VSM and LDA in Fig. 4. Moreover, the fine-grained topics are more distinguishable features than original NMF method, so that the experiment without the coarse-grained topic achieve better result than all feature. In contrast to original NMF, the more features can gain better results.

As for 20 newsgroup datasets, we utilize three methods which are liblinear, KNN and NaiveBayes. Liblinear has a good performance on high-dimensional sparse features in our two classification comparison experiments, so that the contrasts are subtle. As a result of retaining all semantic features, VSM is superior to our model. In contrast to KNN methods, the result shows that the topics we extracted from corpus are distinguishable from others.

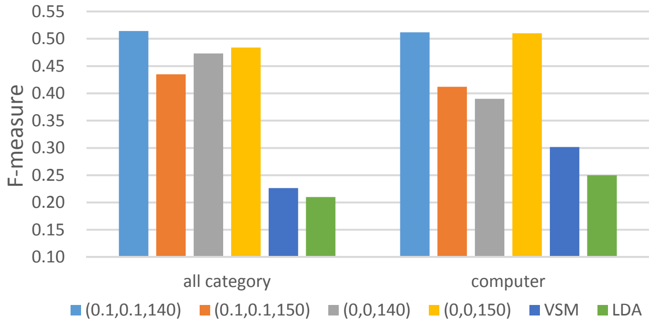

Fig. 4. Clustering preformance of variant parameter of our model and other model (λ_1 , λ_2, k)

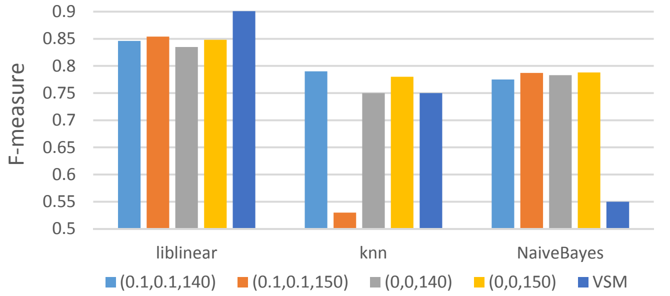

Fig. 5. Classification preformance of various methods in the whole 20newsgroup (λ_1 , λ_2, k)

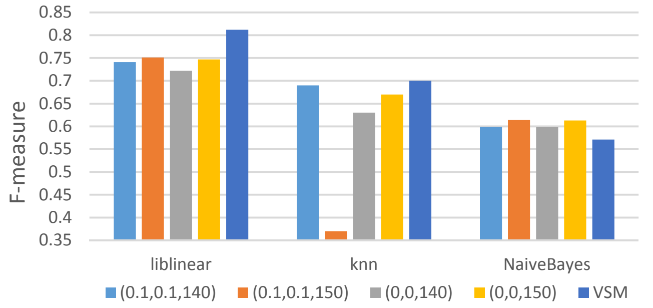

Fig. 6. Classification preformance of various methods in the superclass of computer

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose an improved topic model based on NMF method which characterizes the coarse-topics and fined-topics. This two parts of topics can help us to have a comprehensive understanding of the massive documents. Different people may have various focus points. Specifically, someone who wants to know the overview of the documents consider the coarse-grained topics are valuable, but another one who concerns the details of the same set of documents should pay attention to the fine-grained topics. Our experiments show that the proposed method can tell the differences between coarse-grained topics and find-grained topics, and we proved the coarse-grained topics are really widespread in the given documents. We believe our method can get amazing results in other domain corpus and offer people a new view of data. Future research work will focus on implementing our distributed parallel algorithms on Spark, so that we have more text to train the parameter in the matrix and will achieve better results.

ACKNOWLEDGMENT

Dr. Wang was supported by China Postdoctoral Science Foundation funded project (Grant No. 2014M550591), and Dr. Zhang was supported by the National Natural Science Foundation of China (Grant No. M1450009).

REFERENCES

[1] Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.

[2] Wang, Yu-Xiong, and Yu-Jin Zhang. "Nonnegative matrix factorization: A comprehensive review." Knowledge and Data Engineering, IEEE Transactions on 25.6 (2013): 1336-1353.

[3] Chemudugunta, Chaitanya, and Padhraic Smyth Mark Steyvers. "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model." *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference.* Vol. 19. MIT Press, 2007.

[4] Hofmann, Thomas. "Probabilistic latent semantic indexing." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1999.

[5] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation". the Journal of machine Learning research 3 (2003): 993-1022.

[6] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A.. "Indexing by latent semantic analysis" . JASIS, (1990): 41(6), 391-407

[7] Wang, Q., Xu, J., Li, H., & Craswell, N. "Regularized latent semantic indexing." *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011.

[8] Xu, Z., Chang, X., Xu, F., and Zhang, H.. L1/2regularization: A thresholding representation theory and a fast solver.*IEEE Transactions on neural networks and learning systems*,(2012):23(7), 1013-1027

[9] Newman, D., Smyth, P., Welling, M., and Asuncion, A. U.. Distributed inference for latent dirichlet allocation.InAd-vances in *Neural Information Processing Systems* (2007):pp.1081-1088.

[10] Asuncion, Arthur U., Padhraic Smyth, and Max Welling. Asyn-chronous distributed estimation of topic models for document analysis. *Statistical Methodology* 8.1 (2011): 3-17.

[11] Wold, Svante, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2.1 (1987): 37-52.

[12] Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003.

[13] Gaussier, Eric, and Cyril Goutte. "Relation between PLSA and NMF and implications." *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2005.

[14] Chen, Y., Zhang H., Zuo Y., & Wang D. "An Improved Regularized Latent Semantic Indexing with L1/2 Regularization and Non-negative Constraints." *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on. IEEE*, 2013.

[15] Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections." *Proceedings of the 17th international conference on World Wide Web.* ACM, 2008.

[16] Hoyer, Patrik O. "Non-negative matrix factorization with sparseness constraints." *The Journal of Machine Learning Research 5* (2004): 1457-1469.

[17] Zhai, ChengXiang, Atulya Velivelli, and Bei Yu. "A cross-collection mixture model for comparative text mining." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2004.

[18] Titov, Ivan, and Ryan McDonald. "Modeling online reviews with multi-grain topic models." *Proceedings of the 17th international conference on World Wide Web.* ACM, 2008.