Learning Topics in Short Texts by Non-negative Matrix Factorization on Term Correlation Matrix*

Xiaohui Yan†, Jiafeng Guo†, Shenghua Liu†, Xueqi Cheng†, Yanfeng Wang‡

Abstract

Nowadays, short texts are very prevalent in various web applications, such as microblogs, instant messages. The severe sparsity of short texts hinders existing topic models to learn reliable topics. In this paper, we propose a novel way to tackle this problem. The key idea is to learn topics by exploring term correlation data, rather than the high-dimensional and sparse term occurrence information in documents. Such term correlation data is less sparse and more stable with the increase of the collection size, and can well capture the necessary information for topic learning. To obtain reliable topics from term correlation data, we first introduce a novel way to compute term correlation in short texts by representing each term with its co-occurred terms. Then we formulated the topic learning problem as symmetric non-negative matrix factorization on the term correlation matrix. After learning the topics, we can easily infer the topics of documents. Experimental results on three data sets show that our method provides substantially better performance than the baseline methods.

1 Introduction

Nowadays, short texts are very prevalent in various web applications, such as microblogs, SNS statuses, instant messages, video titles, and so on. These type of data always summarizes all types of information, like the most recent personal information, or news events. Therefore, it is increasingly important to understand and represent short texts properly for many text processing tasks, like emerging topics discovery[4] in social media, efficient index and retrieval[21], personalized recommendation[17], etc.

Topic models, like PLSA[11] and LDA[2], provide a principled way to represent and analyze text collection by uncovering the hidden thematic structure of it automatically. However, conventional topic models usually target at normal text, whose effectiveness will be highly influenced when the document length reduces, as the experiments of [12] shown. An intuitive explanation is like follows. Most conventional topic models, like PLSA and NMF[16], learn topics by decomposing the the so-called term-document matrix into two low-rank matrices, illustrating in Figure 1(a). However, the term-document matrix which represents the term occurrence information in documents, is extremely sparse as for short texts. More formally, Let M, N, K be the number of distinct terms, documents, and topics, respectively. We have $MK + NK$ latent variables to be estimated for the two low-rank matrices U and V, where on average $\frac{MK}{N} + K$ latent variables for each document. When the documents are very short, e.g. with 10 or less terms, the problem becomes highly underdetermined for a large K. In other words, we may not be able to learn the topics reliably for such sparse data.

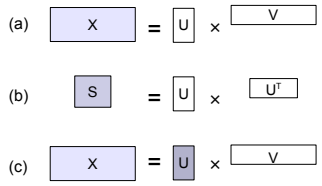

Figure 1: (a) The conventional topic models decompose the extremely sparse term-document matrix X into the term-topic matrix U and the topic-document matrix V; Our approach (b) learns U by symmetric factorization on a dense term correlation matrix S, and then (c) solves the topic-document matrix V with U learned in hand.

To overcome this problem, we propose a novel way to learn topics from short texts. The key idea is that since the term-document matrix is too sparse to estimate reliable topics, we shall turn to more stable and dense data for this purpose. As we know, topics are mainly uncovered based on the correlations between terms. For instance, if the terms "President" and "Obama" co-occur frequently, they might talk about the same topic. Meanwhile, we observe that when the

*This work is funded by the National Natural Science Foundation of China under Grant No. 61202213, No. 60933005, No. 61173008, No. 61003166, and 973 Program of China under Grants No. 2012CB316303.

†Institute of Computing Technology of the Chinese Academy of Sciences, Beijing. Email: yanxiaohui@software.ict.ac.cn, {guojiafeng,liushenghua,cxq}@ict.ac.cn

‡Sogou Inc, Beijing. Email: wangyanfeng@sogou-inc.com

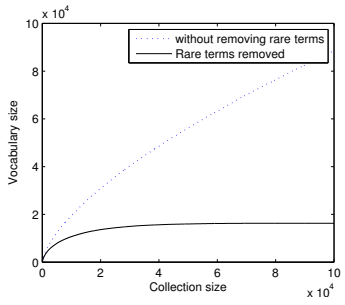

Figure 2: The Vocabulary size grows slowly as the collection size increasing on 10w twitter posts, when rare terms with document frequency < 4 are removed.

size of the short text corpus becomes larger and larger, the size of distinct terms usually keeps relative small and stable. Figure 2 illustrates this phenomenon, as the size of Twitter posts increases from 10^4 to 10^5, the vocabulary size almost keeps the same in Twitter2011 data set, which will be detailed in the Experiments section. Therefore, it is feasible to directly estimate topics from term correlation data rather than the sparse term-document matrix.

The consequent question is how to obtain the term correlation data from short texts. A straightforward way is to directly use the term co-occurrences. It is equivalent to represent each distinct term by a document vector, in which each entry indicates the occurrence of the term in a document, and then measuring correlation between terms via similarity measures like cosine similarity, or Pearson correlation. However, such a way also suffers from the sparsity problem in short texts, since the dimension of document vector could be very high for a large corpus, the term-document vector turns out to be very sparse. Instead, we employ an alternative term correlation measure for short texts by representing each term by a vector of co-occurred terms rather than documents, and then compute correlation of these vectors. Since the vocabulary size is much smaller and more stable than collection size in short texts, this correlation measure does not suffer from the sparsity problem.

Specifically, our approach for topic learning in short texts consists of two steps, shown in Figures 1 (b) and (c). First, we construct a term correlation matrix S based on the proposed term correlation measure. And then we apply symmetric non-negative matrix factorization on the correlation matrix to learn the topics, which result in the term-topic matrix U, as shown in Figure 1(b). Second, we infer the topics of documents by solving the topic-document matrix V, according to the original term-document matrix X and the term-topic matrix U in hand, as shown in Figure 1(c).

We develop efficient algorithms for topic learning and inference in short texts, and test our approach on three real-world short text data sets. Experimental results demonstrate the effectiveness of the proposed topic learning method in topic visualization, document clustering, and document classification.

The major contributions of our approach on short text topic learning are as follows:

- So far as we know, it is the first attempt to learn topics directly from the term correlation matrix rather then term-document matrix. In this way, we can largely alleviate the data sparsity problem when applying topic models on large scale short text collections.
- We propose to measure the correlation between terms by representing each term with its co-occurred terms rather than the documents it occurred. Hence, we can avoid the sparsity problem in term representation and better measure the term correlation.
- We developed efficient algorithms for topic learning and inference in our approach.
- We conduct extensive experiments to demonstrate the effectiveness and efficiency of the proposed approach.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 describes our solution to short texts topic model. Section 4 discusses the algorithms for topic learning and inference. Experimental results are presented in Section 5. Conclusions are made in the last section.

2 Related Work

There is considerable amount of literature on topic learning for text data in the past decade. From the view of methodology, they fall into two groups: non-probabilistic approaches and probabilistic approaches.

Most of non-probabilistic approaches are based on matrix factorization techniques, which project the term-document matrix into a K-dimensional topic space. For example, the early work LSI [6] employed SVD to identify latent semantic factors with orthogonal constraints. Some recent works utilized the sparse constraint to substitute the orthogonal ones, like the regularized LSI [22] and sparse LSA [5]. However, these methods lack an intuitive interpretation for the negative values in results [24]. A more popular way is NMF [16, 24, 19], which

introduced non-negative constraint instead. With the non-negative constraints, only additive operator is allowed in factorization. Therefore, NMF is believed to learn part-based representations of dataset. From the view of topic modeling, NMF decomposes the term-document matrix into two low-rank non-negative matrices: a *term-topic matrix*, each column represents a topic as a convex combination of terms; and a *topic-document matrix*, each column represents a document as a convex combination of topics. Our work falls into this group.

Probabilistic topic models are also very popular. The representative models are the probabilistic Latent Semantic Analysis (PLSA) of Hofmann [11], and the Latent Dirichlet Allocation (LDA) of Blei[2]. PLSA models each document as a mixture of topics, which equals to NMF with KL divergence[10]. LDA incorporates the Dirichlet priors for topic mixtures, thus it is less prone to over-fitting and capable of inferring topics for unobserved documents.

While all of the above topic models deal with normal texts, there are some recent works advocating for short text medias, such as Twitter posts. For example, [23, 12] proposed to train LDA on "fake" documents by aggregating tweets of users. Ramage et.al[18] developed a partially supervised learning method (Labeled LDA) to model Twitter posts with the help of hashtag labels. Yan et.al[25] proposed the Ncut-weighted NMF for short text clustering by utilizing term correlation information, too. Different from them, we studied the problem of topic learning for general short texts which is domain-independent.

3 Our Approach

In our work, we learn topics from term correlations, rather than term-documents matrix, which is extremely sparse for short texts. At first, a term correlation matrix is constructed. And then, symmetric non-negative matrix factorization is performed to learn the topic representation of the terms. Finally, by projecting documents into the low-dimensional topic space, we are capable to infer the topical representation of each document.

3.1 Term Correlation Matrix Construction

Conventionally, documents are presented via the well-known Vector Space Model, which models a collection of documents by a term-document matrix X, with each entry x_{ij} indicating the weight (typically by TFIDF) of the term t_i in document d_j. In other words, a term t_i can be viewed as a document vector $(x_{i,1}, ..., x_{i,N})$, where N is the number of documents in the collection. To measure the correlation of two terms, a common way is to calculate the similarity of their vector representations. However, the document vector representation still suffers from the extreme sparsity problem in short text data, because of the shortness of short texts limits each term occurs in a relative small part of documents in a collection. Figure 3(a) visualizes the term correlation matrix computed via document vector represenation on the Tweets corpus. We can see the result matrix is highly sparse, only with denstiy 0.0415.

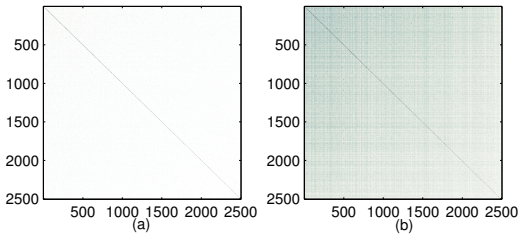

Figure 3: Visualization of Term correlation matrix on Tweets corpus computed via (a) document vector representation of terms, density=0.0415; (b) via co-occurred term vector representation of terms, density=0.8835.

Instead, here we suggest to represent a term by a term co-occurrence vector, rather than the document vector. The original idea comes from the famous dictum—"You shall know a word by the company it keeps!"[9], in natural language processing field, which tells us that the meaning of a word can be decided by the distribution of words around it. For example, considering the word "xbox" occurring in the following short text snippets: "xbox live game downloading", "can xbox play Left 4 Dead", and "xbox 360 vs. ps3", it is not difficult to deduce that "xbox" refers to a game machine on the basis of the co-occurred words "game", "play" and "ps3".

After then, we can also measure the similarity of two terms according to their co-occurred words distribution. In detail, a term t_i is presented by a term occurrence vector $(w_{i,1}, ..., w_{i,M})$, where $w_{i,m}$ is decided by the co-occurrence of terms t_i and t_m. In this work, we apply the positive point mutual information (PPMI) to assess $w_{i,m}$:

$$w_{i,m} = \text{PPMI}(t_i, t_m) = \max(\log \frac{P(t_i, t_m)}{P(t_i)P(t_m)}, 0),$$

where the probabilities $P(t_i, t_m)$ and $P(t_i)$ are estimated empirically:

$$P(t_i, t_m) = \frac{\text{n}(t_i, t_m)}{\sum_{j,l} \text{n}(t_j, t_l)}, P(t_i) = \frac{\sum_m \text{n}(t_i, t_m)}{\sum_{j,l} \text{n}(t_j, t_l)},$$

where $n(t_i, t_m)$ is the times of terms t_i and t_m co-occurred. After representing each term by the term co-

occurrence vector, we then apply the common vector-similarity measures, like cosine coeffcient, to compute the correlation between any two terms, resulting in the final term correlation matrix S.

The above method to estimate the correlation between two terms is known as distributional methods in natural language processing fields[14]. Figure 3(b) visualizes the term correlation matrix computed via term co-occurrence vector representation on the Tweets corpus, it is with much higher density(0.8835) than Figure 3(a).

3.2 Topics Learning In probabilistic topic models, a "topic" is considered as a distribution over terms[2]. Moreover, a meaningful "topic" should be a distribution concentrated on terms about a particular subject. Accordingly, in non-probabilistic topic models, a "topic" can be viewed as a group of terms with weights indicating the importance or significance in some subject. Consequently, to discover the topics equal to cluster the terms into some meaningful groups. On the other hand, terms usually have multiple meanings, and may belong to multiple topics. Therefore, a direct way to perform topic learning is to conduct "soft" clustering based on the term correlation matrix.

Motivated by previous works about graph clustering[15], we formulate this topic learning problem as finding a term-topic matrix U to minimize the following objective function:

$$L(U) = \|S - UU^T\|_F^2, \quad s.t.\ U \geq 0., \tag{3.1}$$

where each column of the term-topic matrix U represents a topic by a vector of weighted terms. This special formulation of non-negative matrix factorization is referred as the symmetric non-negative matrix factorization, which is suggested to be equivalent to kernel Kmeans clustering and spectral clustering[7].

3.3 Topics Inference for Documents The term-topic matrix U uncovers the latent topic structure of the collection. Once it obtained, we can subsequently infer the topic presentations of documents, namely the topic-document matrix V by projecting the documents into the latent topic space. In the non-negative matrix factorization framework, this problem can be formulated as finding a non-negative matrix V to minimize the distance between the product UV and X, while given term-topic matrix U.

To judge the optimal solution, we need to know the how well the estimated model fits the observed data. Two popular distance functions are used for measuring the lost. One is the square of the Euclidean distance:

$$L_E(V) = \| X - UV \|_F^2 . \tag{3.2}$$

The other one is the generalized I-divergence as the following:

$$\begin{aligned} L_I(V) &= D(X \parallel UV) \\ &= \sum_{ij} \left(x_{ij} \log \frac{x_{ij}}{(UV)_{ij}} - x_{ij} + (UV)_{ij} \right), \end{aligned} \tag{3.3}$$

which reduces to the Kullback-Leibler divergence if $\sum_{ij} x_{ij} = \sum_{ij} (UV)_{ij} = 1$.

In both Eq. (3.2) and Eq. (3.3), we can find that each column $\mathbf{v}_i$ in V only dependents on the column $\mathbf{x}_i$ in X, suggesting that we can infer the topics of each document independently. Hence, it is easy to parallelize the topic inference procedure for large-scale data.

The major difference between our approach and the standard NMF is that we separate the topic modeling process in NMF into two sub-tasks: topic learning and topic inference for documents, which are proceeded sequentially. In the topic learning stage, we solves a smaller and denser matrix factorization problem in Eq. (3.1). While in the topic inference stage, we solves a non-negative least squares problem in Eq. (3.2) or Eq. (3.3). Both of the two sub-task is much easier to solve than directly factorizing the extremely sparse and large term-document matrix. For convenience, we denote our approach as "TNMF" in the following description.

4 Learning Algorithm

In this section, we will detail the algorithms to learn the term-topic matrix U and the topic-document matrix V, respectively.

Algorithm 1: The overall procedure of our approach

Input: the topic number K, the term-document matrix $X \in R^{M \times N}$

Output: the term-topic matrix $U \in R^{M \times K}$, and the topic-document matrix $V \in R^{K \times N}$

compute term correlation matrix $S \in R^{M \times M}$
random initialize U
repeat
  $U \leftarrow \max(SU(U^TU)^{-1}, 0)$
until *convergence*;
$V = \max((U^TU)^{-1}U^TY, 0)$
return U, V

4.1 Solving the Term-Topic Matrix U Different from the standard NMF, the objection function of symmetric non-negative matrix factorization in Eq.(3.1) is quartic non-convex function. However, we can access it via conventional NMF solver by treating U and

U^T as two different matrices, and then update them alternatively as follows:

$$U_{t+1} = \underset{U \geq 0}{\operatorname{argmin}} \parallel S - U_t U^T \parallel^2, \tag{4.4}$$

where U_t is the term-topic matrix solved in tth iteration. When this update converge to a stationary point, we have $U_{t+1} = U_t$. It is easy to see the stationary point is a solution of Eq.(3.1).

Eq.(4.4) is a non-negative least squares (NNLS) problem. If U_t has full rank, the objective function of the NNLS problem is strictly convex, which is guaranteed to have a unique optimal solution. However, exactly solving NNLS is much slower than solving unconstrained least squares problems, especially in high dimensional space[1]. Instead, we try to find an approximate solution by enforcing the non-negativity constraint by setting all the negative elements resulting from the least squares solution to 0, namely:

$$U_{t+1} = \max(SU_t(U_t^T U_t)^{-1}, 0). \tag{4.5}$$

Although lack of convergence theory for this ad-hoc enforcement, this approximation algorithm often converge quickly and give very accurate results in practice[1].

4.2 Solving the Topic-Document Matrix V After learning matrix U, we then solve the topic-document matrix V. If the Euclidean distance used, minimizing the objective function (3.2) is also a NNLS problem. As the same as solving matrix U, we can obtain the following update rule:

$$V = \max((U^T U)^{-1} U^T Y, 0). \tag{4.6}$$

If the generalized I-divergence used, the objective function (3.3) can be write as minimizing N independent sub-problems as follows:

$$l(\mathbf{v}_j) = \sum_{i=1}^{M} \left(x_{ij} \log \frac{x_{ij}}{\sum_{k=1}^{K} u_{ik} v_{kj}} - x_{ij} + \sum_{k=1}^{K} u_{ik} v_{kj} \right)$$

Unfortunately, this problem has no closed form solution. We then apply the coordinate descent method to solve it. For each variable v_{kj}, we fix all other variables $v_{k'j}(k' \neq k)$ in vector $\mathbf{v}_j$ and apply Newton's method to update v_{kj}. The first order derivative $g(v_{kj})$ and the second order derivative $h(v_{kj})$ of l on v_{kj} is

$$g(v_{kj}) = \frac{\partial l}{\partial v_{kj}} = \sum_{i=1}^{M} u_{ik} \left(1 - \frac{x_{ij}}{\sum_{k=1}^{K} u_{ik} v_{kj}} \right),$$

$$h(v_{kj}) = \frac{\partial^2 l}{\partial v_{kj}^2} = \sum_{i=1}^{M} \frac{x_{ij} u_{ik}^2}{(\sum_{k=1}^{K} u_{ik} v_{kj})_i^2}.$$

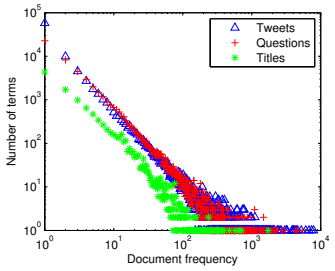

Figure 4: Distribution of terms in three short text data sets

And the Newton update rule is:

$$v_{kj} \leftarrow \max(v_{kj} - \frac{g(v_{kj})}{h(v_{kj})}, 0).$$

Algorithm 1 summarizes the overall procedure of our approach using the Euclidean distance for illustration.

5 Experiments

In this section, we report empirical experiments on real-world short text data sets to demonstrate the effectiveness of our approach.

5.1 Data Sets We carried out experiments on three data sets. 1) Tweet data, a subset of TREC 2011 microblog track [1]. 2) Title data, including news titles with class labels from some news websites, which is published by Sogou Lab [2]. 3) Question data, containing questions crawled from a popular Chinese question-and-answer website[3]. Each question is with a manual class label assigned by the questioner.

Figure 4 shows the distributions of terms in the three sets of short text data. We can see that there are a large proportion of terms occur in less than 10 documents in all the three data sets. With further investigation, we find more than half of those rare terms are meaningless, e.g. "witit","25c3", etc. Therefore, we preprocessed the raw data by removing words with document frequency less than 10.

The data sets after preprocessing are summarized in Table 1. In the two relative small data sets, i.e. the Tweet data and the Title data, the number of documents is only about twice of the number of distinct terms. But in the much larger Question data, the number of documents is more than 7 times of distinct terms.

[1] http://trec.nist.gov/data/tweets/
[2] http://www.sogou.com/labs/dl/tce.html
[3] http://zhidao.baidu.com

Table 1: Statistics of the three data sets

Data sets	Tweet	Title	Question
#documents	4520	2630	36219
#words	2502	1403	4956
avg words†	8.5958	5.2684	5.8092
#classes	unavailable	9	34

† denotes the average number of words in a document

5.2 Baselines Our baseline methods include:

- LDA. We used the Gibbs sampling based LDA implementation GibbsLDA++[4].
- NMF. We compared with NMF with two different cost functions: "NMF_E" denotes the NMF with the Euclidean distance based cost function; And "NMF_I" denotes the NMF with the generalized I-divergence. Note we added the ℓ_2 norm regularizer for "NMF_E" to avoid over-fitting as [19] did. However, "NMF_I" is without any regularization.
- GNMF(graph regularized NMF)[3]. GNMF directly factorizes the term-document matrix, while employing a Laplace regularization to keep the local neighboring relationship. The neighborhood relationship of documents is generated by the cosine similarity measure. Here, we use the authors' code [5].
- SymNMF(symmetric NMF)[15]. SymNMF factorizes a symmetric matrix containing pairwise document similarity values, measured by cosine similarity, too.

All the parameters of baseline methods were tuned to best manually[6]. In addition, we use "TNMF_E" to denote our TNMF with Euclidean distance in learning document-term matrix V, and use "TNMF_I" to denote the generalized I-divergence based counterpart.

5.3 Topic Visualization Interpretability of topic models is very important. In most applications, we prefer topic models generating topics with good readability and compact representation. For this reason, we compared the hidden topics discovered by all the test methods qualitatively. Considering the topics learned by different methods are very different, we only select some similar topics among them for comparison.

[4]http://gibbslda.sourceforge.net

[5]http://www.cad.zju.edu.cn/home/dengcai/Data/GNMF.html

[6]In GNMF, we find the best value of the regularization coefficient sometimes is 0, which reduces to the standard NMF. In such case, we set it to 0.01 for the purpose of comparison.

Table 2 shows 4 topics from Tweet data (K=20), and each topic represented by its top 5 weighted terms. It is clear that they talk about "Egyptian revolution", "business", "weather", and "Super Bowl", respectively, which are hot in the Tweet data. Noting that the original SymNMF does not learn the term-topic matrix, hence we did not compare with it. From Table 2 we can conclude that

- LDA is able to discover meaningful topics, but it assigns high weights to some common terms, such as "medium" and "change".
- NMF_E generated topics more discriminative than LDA, for the sake of IDF weighting used to improve term's discriminative power. However, some topics still need effort for explanations, e.g. the "Super Bowl" and "business" topics.
- NMF_I failed in finding meaningful topics. The possible reason is over-fitting, since NMF_I does not employ any regularization.
- GNMF showed similar results with NMF_E, suggesting that the document neighborhood regularization plays little effect on short texts.
- TNMF_E and TNMF_I generated much more compact topics with the best readability.

5.4 Document Clustering Document clustering is an important technique to automatically group similar documents in corpus, and widely used in various applications, such as navigation, retrieval, and summarization of huge volumes of text documents. Topic models also have the effect in grouping similar documents, but they allow documents belonging to multiple groups. To utilize topic models for document clustering, we assigned each document to the highest weighted topic.

5.4.1 Evaluation Metrics Assuming $\Omega = \{\omega_1, ..., \omega_K\}$ is the set of clusters, each ω_k is the document set in cluster k. And $\mathbb{C} = \{c_1, ..., c_P\}$ is the set of P classes of test documents labeled ahead as the ground truth. We adopt three standard criteria to measure the quality of clusters set Ω:

- Purity[26]. Supposing documents in each cluster should take the dominant class in it, purity is the accuracy of this assignment measured by counting the number of correctly assigned documents and dividing by the total number of test doucments. Formally:

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{n} \sum_{i=1}^{k} \max_{j} |\omega_i \cap c_j|$$

Table 2: Top words in 4 topics discovered by different methods on the Tweet data.

Topic	LDA	NMF_E	NMF_I	GNMF	TNMF_E	TNMF_I
Egyptian revolution	egypt	egyptian	house	egypt	egyptian	egyptian
	food	egypt	egypt	state	cairo	cairo
	state	mubarak	update	egyptian	protester	protester
	egyptian	cairo	state	president	mubarak	egypt
	report	protester	100	report	egypt	mubarak
business	service	market	market	market	market	market
	medium	business	red	business	debt	company
	stand	social	white	social	credit	financial
	market	medium	hey	medium	company	business
	job	company	die	online	business	finance
weather	hot	wind	snow	wind	humidity	humidity
	wind	humidity	fall	humidity	temperature	temperature
	change	temperature	wind	rain	mph	mph
	fall	rain	street	temperature	hpa	hpa
	humidity	mph	humidity	mph	pressure	wind
Super Bowl	super	super	super	super	packer	packer
	bowl	bowl	bowl	bowl	nfl	nfl
	green	xlv	green	xlv	bay	bay
	team	sunday	red	packer	bowl	bowl
	fan	party	heart	sunday	rodgers	rodgers

Note that when documents in each cluster are with the same class label, purity is highest with value of 1. Conversely, it is close to 0 for bad clustering.

- Normalized Mutual Information(NMI)[20]. NMI measures the mutual information $I(\Omega;\mathbb{C})$ penalized by the entropies $H(\Omega)$ and $H(\mathbb{C})$:

$$\begin{aligned}\mathrm{NMI}(\Omega,\mathbb{C}) &= \frac{I(\Omega;\mathbb{C})}{[H(\Omega)+H(\mathbb{C})]/2} \\ &= \frac{\sum_{i,j}\frac{|\omega_i\cap c_j|}{n}\log\frac{|\omega_i||c_j|}{n|\omega_i\cap c_j|}}{(\sum_i\frac{|\omega_i|}{n}\log\frac{|\omega_i|}{n}+\sum_j\frac{|c_j|}{n}\log\frac{|c_j|}{n})/2}\end{aligned}$$

NMI is 1 for perfect match of Ω and $\mathbb{C}$, and 0 when the clusters are random with respect to the class membership.

- Adjusted Rand Index(ARI)[13]. The Rand Index measures the accuracy of pair-wise decisions, i.e. the ratio of pairs of objects which are both located in the same cluster and the same class, or both in different clusters and different classes. Adjusted Rand Index is the corrected-for-chance version of the Rand Index, which yield a value between [-1, 1]. The higher the Adjusted Rand Index, the more resemblance between the clustering results and the labels.

$$ARI = \frac{\sum_{i,j}\binom{|\omega_i\cap c_j|}{2} - [\sum_i\binom{|\omega_i|}{2}\sum_j\binom{|c_j|}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i\binom{|\omega_i|}{2}+\sum_j\binom{|c_j|}{2}] - [\sum_i\binom{|\omega_i|}{2}\sum_j\binom{|c_j|}{2}]/\binom{n}{2}}$$

5.4.2 Quantitative Evaluation Quantitative evaluation is conducted on the two data sets with label information: the Title data and the Question data, with the number of clusters ranging from 20 to 100. We have run 10 times for each evaluation, and the average performance scores are reported.

Figure 5 and Figure 6 illustrate the comparison of each method's clustering performance on the two data sets. From the results, we can draw the following conclusions.

- Overall, TNMF_E and TNMF_I always outperform all the baseline methods in all three evaluation metrics, especially in terms of NMI and ARI. In the Question data, the improvement is much more significant than in the Title data. It demonstrates that topics can be accurately identified from the term correlation matrix.

- Among the baseline methods, NMF_E is slightly worse than LDA in most cases, but comparable with or even better than GNMF and SymNMF. It suggests that the document similarity information do not benefit the clustering performance in these short texts.

- Besides, NMF_KL performed very bad, and the possible reason is that the sparse data make it overfit the data.

- Surprisingly, the results show that SymNMF and TNMF are not sensitive to over-fitting. The pos-

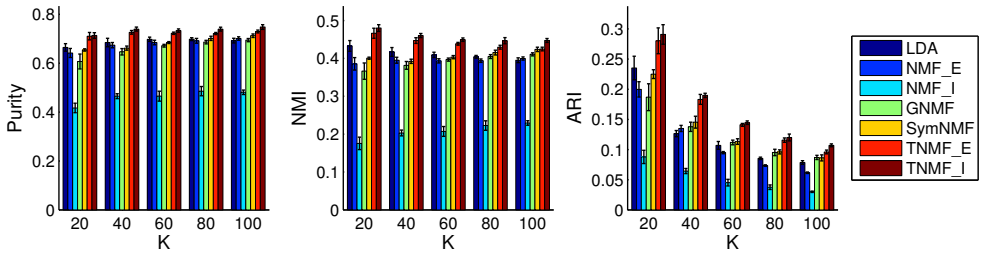

Figure 5: Comparison of (a)Purity, (b)NMI, (c)ARI w.r.t the cluster number k on the Title data

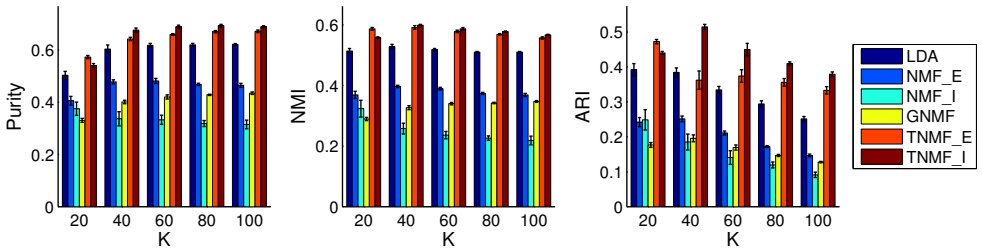

Figure 6: Comparison of (a)Purity, (b)NMI, (c)ARI w.r.t the cluster number k on Question data

sible reason is that the symmetric matrix factorization has only one sub-matrix to be estimated. The freedom of latent variables are much less than NMF.

5.5 Document Classification To verify how well the documents are represented by the learned topics, we further tested the classification accuracy of short texts by representing documents in the latent spaces.

The document classification experiments were conducted on both the Title data and the Question data, too. In each data set, documents are randomly split into training and testing sub-sets with the ratio 4 : 1, then classified by the linear support vector machine classifier LIBLINEAR[8].

Figure 7 and Figure 8 show the classification results for all the test methods with topic number K varying from 20 to 100 on the two data sets. These experiments reveal several interesting points:

- TNMF_E substantially outperforms baseline methods on both data sets, especially in the Question data which has much more documents. It suggests that TNMF_E can capture the topics of documents more accurately than other methods.

- The performance of TNMF_I is not as good as TNMF_E in classification, but still better than the baseline methods on the Question data, and comparable with them on the Title data.

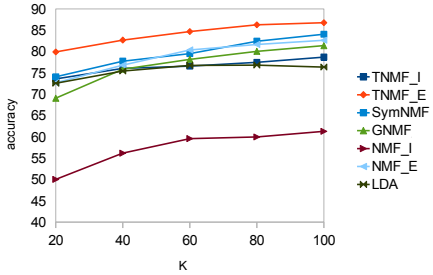

Figure 7: Classification accuracy on the Title data

- GNMF and SymNMF do not achieve any improvement over NMF on the Title data, while even decline in terms of accuracy on the Question data. It implies that the document neighborhood relationship might not be accurate for short texts any more.

6 Conclusions

Learning topics for short texts is considered to be a difficult problem due to the severe sparsity of term-document co-occurrence data. To tackle this problem, we have presented a novel method based on the non-negative matrix factorization framework. Our approach first learns topics from term correlation data using symmetric non-negative matrix factorization, and then infers the topic representations of documents by solving a non-negative non-negative least squares problem. Since

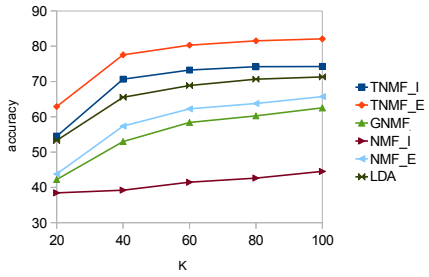

Figure 8: Classification accuracy on the Question data

the term correlation data is less sparse and more stable with the increase of the collection size, our approach is able to learn better topics than tranditional methods. Experiments on three short text data sets illustrated superior performance of our methods compared with the other baseline methods.

References

[1] R. Albright, J. Cox, D. Duling, A. Langville, and C. Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, 2006.

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

[3] D. Cai, X. He, J. Han, and T.S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[4] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[5] X. Chen, B. Bai, Y. Qi, Q. Lin, and J. Carbonell. Sparse latent semantic analysis. In *NIPS Workshop*, 2010.

[6] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[7] C. Ding, X. He, and H.D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proc. SIAM Data Mining Conf*, pages 606–610, 2005.

[8] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[9] John R. Firth. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32, 1957.

[10] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 601–602. ACM, 2005.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM, 1999.

[12] L. Hong and B.D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[13] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[14] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition, February 2008.

[15] D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proc. SIAM Data Mining Conf*, 2012.

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

[17] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, pages 385–388, 2009.

[18] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2010.

[19] F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[20] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

[21] J. Teevan, D. Ramage, and M.R. Morris. # twittersearch: a comparison of microblog search and web search. In *WSDM*, pages 35–44. ACM, 2011.

[22] Q. Wang, J. Xu, H. Li, and N. Craswell. Regularized latent semantic indexing. In *SIGIR*, pages 685–694. ACM, 2011.

[23] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, New York, NY, USA, 2010. ACM.

[24] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273. ACM, 2003.

[25] Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. Clustering short text using ncut-weighted non-negative matrix factorization. In *CIKM*, pages 2259–2262, 2012.

[26] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. 2001.