Clustering Short Text Using Ncut-weighted Non-negative Matrix Factorization

Xiaohui Yan, Jiafeng Guo
Institute of Computing Technology, CAS
Beijing, China 100190
{yanxiaohui, guojiafeng@software.ict.ac.cn}

Shenghua Liu*, Xue-qi Cheng
Institute of Computing Technology, CAS
Beijing, China 100190
{liushenghua, cxq}@ict.ac.cn

Yanfeng Wang
Sogou Inc.
Beijing, China 100084
wangyanfeng@sogou-inc.com

ABSTRACT

Non-negative matrix factorization (NMF) has been successfully applied in document clustering. However, experiments on short texts, such as microblogs, Q&A documents and news titles, suggest unsatisfactory performance of NMF. An major reason is that the traditional term weighting schemes, like binary weight and *tfidf*, cannot well capture the terms' discriminative power and importance in short texts, due to the sparsity of data. To tackle this problem, we proposed a novel term weighting scheme for NMF, derived from the *Normalized Cut* (Ncut) problem on the term affinity graph. Different from *idf*, which emphasizes discriminability on document level, the Ncut weighting measures terms' discriminability on term level. Experiments on two data sets show our weighting scheme significantly boosts NMF's performance on short text clustering.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval; I.5.3 [**Pattern Recognition**]: Clustering

Keywords

Short Text, Clustering, NMF, Normalized Cut

1. INTRODUCTION

Short texts are prevalent on the web nowadays, such as microblogs, SNS statuses, and instant messages, etc. They are with limited document length, typically only tens of words or even less on average. Successfully clustering these data is very important in many web applications, e.g. emerging topics discovery, efficient index and retrieval, and personalized recommendation. However, traditional document clustering methods cannot accomplish the task effectively due to the insufficient representation of short texts.

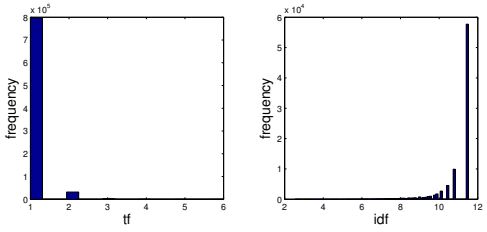

Figure 1: frequency of (a) *tf* values, (b) *idf* values of terms in the Questions data.

One crucial question, for conventional clustering methods like Kmeans and NMF[5], is how to weight terms in such short documents. For normal texts, the widely used weighting method is *tfidf*, defined as follows:

$$tfidf_{t,d} = tf_{t,d} \times idf_t$$
$$idf_t = \log \frac{N}{df_t},$$

where $tf_{t,d}$ is the term frequency of term t in document d, measuring the importance of t in d; df_t is the document frequency of term t in corpus, measuring the discriminative power of t over the entire corpus; and N is the total number of documents in corpus. However, in short text both tf and idf are not very differentiable. Figure 1(a) shows the distribution of tf values in the Tweets data, suggesting that about 96% of them with equal to 1. In other words, most of terms usually occur only once in a short document. Figure 1(b) shows the distribution of idf values, which is dominated by values larger than 8. What is even worse is that about 65% of terms have the same idf value 11.51—the highest one of all.

Both of the two problems of *tfidf* come from the insufficient representation of documents, indicating it is not a good idea to weight terms with only document level information. Instead, we propose a novel term weighting scheme for NMF on short text clustering based on words co-occurrence information. This weight is derived from the Normalized Cut (Ncut) problem[3] on term affinity graph, referred as Ncut-weight of terms (Section 2). We show the detail of the Ncut-weighted NMF solution with the Alternating Non-negative Least Squares (ALNS) algorithm (section 3). Both quali-

*corresponding author

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

tative and quantitative evaluations were conducted on two short text data sets: tweets, and web page titles (Section 4). The results demonstrate the effectiveness of the Ncut-weighted NMF.

2. NCUT ON TERM AFFINITY GRAPH

NMF clusters documents and words simultaneously. However, we firstly consider the sub-problem of words clsutering alone, which can be formulated as a graph cut problem. Then, we derive the new term weighting scheme by bridging Ncut on predifined term affinity graph and the traditional NMF method on term-document matrix.

Notations: Let M be the number of distinct terms, N be the number of documents, K be the number of clusters. $X = \{x_{ij}\} \in R^{M\times N}$ denotes the term-document matrix[1]. Moreover, let $\mathbf{x}_i$ denote the ith column vector of X, $\mathbf{x}^{(j)}$ denote the jth row vector of X.

2.1 Term Clustering by Ncut

It is known that words semantic relations can be induced from their co-occurrence frequency. The basic assumption is that if words co-occur frequently, they are likely to semantically related. Based on this assumption, we constructed a term affinity graph $G = \{V, E\}$ to model term similarity according to their co-occurrences. In which, V is given by the term set in corpus, while edge set E determined by predefined adjacent matrix $S = \{s_{ij}\} \in R^{M\times M}$. For example, while inner product is used to measure the similarity,

$$S = XX^T. \tag{1}$$

It is easy to see $s_{ij} = \mathbf{x}^{(i)}\mathbf{x}^{(j)T}$ equals to the number of co-occurrences of term i and term j. If each $\mathbf{x}^{(i)}$ is normalized with unit length, s_{ij} is the cosine similarity.

Clustering terms is equivalents to cut graph G into K sub-graphs. A typical criterion to do that is called the normalized cut criterion[3] that minimizes the normalized weight summation of edges between these sub-graphs. Let $\{G_k\}_{k=1,..,K}$ be a partition, sub-graph $\overline{G_k}$ be the complement of sub-graph G_k, and $S(G_k, G_{k'}) = \sum_{i\in G_k}\sum_{j\in G_{k'}} s_{ij}$, i.e. the weight summation of edges between sub-graph G_k and $G_{k'}$. Thus, the normalized cut problem aims to minimizing the following discrete objective function:

$$Ncut(G_1, \ldots, G_k) = \frac{1}{2}\sum_{k=1}^{K}\frac{S(G_k, \overline{G_k})}{S(G_k, G)}, \tag{2}$$

where $S(G_k, G) = S(G_k, G_k) + S(G_k, \overline{G_k})$.

2.2 Connection between Ncut and NMF

Let $D \in R^{M\times M}$ be the diagonal degree matrix of S, with non-zero entries $d_{ii} = \sum_{j=1}^{M} s_{ij}$, and we define an indicator matrix $U \in R^{M\times K}$, with each entry u_{ik} indicates whether term i belongs to sub-graph G_k:

$$u_{ik} = \begin{cases} \frac{\sqrt{d_{ii}}}{\sqrt{S(G_k,G)}} & t_i \in G_k \\ 0 & otherwise \end{cases}. \tag{3}$$

[1]There are various various schemes to determine x_{ij}, such as *tf* or *tfidf*. In this paper, we takes x_{ij} as binary weight, i.e. if term i occurs in document j, $x_{ij} = 1$, otherwise 0.

It has been proven that the normalized cut criterion can be represented by the following trace maximization problem[6]:

$$\max_U Tr(U^T D^{-1/2} S D^{-1/2} U), \tag{4}$$

where U subjects to constraint (3), which has $U^T U = I$.

Directly tackling the problem of (4) is NP-hard. However, we will show that the approximation solution to this problem can be obtained by NMF with the appropriate weighting scheme.

THEOREM 1. *Non-negative factorization on matrix $Y = D^{-1/2}X$ equals to solving (4) with the discrete constraint Eq. (3) relaxed.*

PROOF. Let $\|\cdot\|_F$ denote the Frobenius norm. The objective function of NMF on Y can be written as

$$\begin{aligned} J(U,V) &= \| Y - UV \|_F^2 \\ &= Tr(Y^TY - 2Y^TUV + V^TU^TUV). \end{aligned}$$

By set the gradient of J over V to zero:

$$\frac{\partial J}{\partial V} = -2U^TY + 2U^TUV = 0.$$

If U^TU is non-singular[2], we get $V = (U^TU)^{-1}U^TY$. Substitute it to J and discard constants, minimization of J is equivalent to

$$\max_U Tr(Y^TU(U^TU)^{-1}U^TY).$$

If with the constraint $U^TU = I$, we have

$$\begin{aligned} \max_U Tr(Y^TUU^TY) &= \max_U Tr(U^TYY^TU) \\ &= \max_U Tr(U^TD^{-1/2}SD^{-1/2}U). \end{aligned}$$

When $U \geq 0$ and $U^TU = I$, $V = U^TY \geq 0$. Therefore, NMF on Y solves problem (4) with the discrete constraint Eq. (3) relaxed. □

3. NCUT-WEIGHTED NMF

In previous section, we derive a term weighting matrix $D^{-1/2}$, i.e. the weight of term i is

$$w_i = d_{ii}^{-1/2} = \left(\sum_{j=1}^{M} s_{i,j}\right)^{-1/2} \tag{5}$$

In practice, it is better to scale all weights into $[0, 1]$, by simply dividing each w_i by $max(\{w_i\}_{i=1,\ldots,M})$. We have to stress that our Ncut-weight is a term weight scheme. [5] uses a similar weight, but actually it is document weight.

From Eq. (5), we can see that if a term co-occurs more frequently with more others, its Ncut-weight will be lower. Comparing with *idf*, the Ncut-weight favors terms with low co-occurrence frequency, rather than document frequency. That is because the words co-occuring frequently tend to to be meaningless or polysemous words, which is not discriminative for clustering. Besides, the term co-occurrence frequency is not highly depend on the document length as document frequency does.

[2]It is always true, since $K \ll M$.

For Ncut-weighted NMF, we also add ℓ_2 norm regularizer to avoid overfitting for such sparse data. The overall object function of Ncut-weighted NMF is

$$\begin{array}{ll} J(U,V) & =\| Y - UV \|_F^2 + \lambda(\| U \|_F^2 + \| V \|_F^2) \\ s.t. & U \geq 0, V \geq 0. \end{array} \quad (6)$$

We employ the alternating non-negative least squares(ANLS) algorithm [1] to solve it, as shown in Algorithm 1.

Algorithm 1: The ANLS algorithm for Ncut-weighted NMF

Input: the number of clusters K, regularization parameter λ, term-document matrix $X \in R^{M \times N}$

Output: $U \in R^{M \times K}$, $V \in R^{K \times N}$

Compute matrices S, D and $Y = D^{-1/2}X$
Initialize U with columns randomly selected in Y
repeat
  $V \leftarrow \max((U^T U + \lambda I)^{-1} U^T Y, 0)$
  $U \leftarrow \max((V^T (V V^T + \lambda I)^{-1}, 0)$
until *convergence*;

4. EXPERIMENTS

4.1 Experiments Setting

4.1.1 Data sets

We carried out experiments on two data sets. 1) Tweets data, collected from twitter.com. 2) Titles data, news titles with assigned class labels from some news websites, which is published by Sogou Lab[3].

The raw data is preprocessed via the following steps: 1) stemming and removing stop words; 2) removing words with document frequency less than 6; 3) removing documents containing less than 4 words. The data characteristics after preprocessing are summarized in Table 1.

Table 1: Description of the data sets

Data sets	#doc	#word	avg words†	#class
Tweets	4520	2502	8.5958	unavailable
Titles	2630	1403	5.2684	9

† denotes average words in a document

4.1.2 Baseline Methods

Our baseline methods include:

- **Kmeans** with terms weighted by *idf*.
- **RLSI** (regularized latent semantic indexing)[4] is a recently proposed method to indentify topics in documents by exploring matrix factorization too. The difference between NMF is RLSI introduces sparse constrains, rather than non-negative ones. Since RLSI has the similar formulation with NMF, we also compared both idf (referred to "RLSI+idf") and Ncut-weights (referred to "RLSI+nc") with it.
- **NMF** with three type term weights: "NMF+01" uses binary weight; "NMF+idf" uses *idf* weight; "NMF+nc" uses Ncut-weight. All of them are regularized by the ℓ_2 norm defined in Eq. (6) with $\lambda = 1$. In our experiments, different λs in $(0.1, 10)$ cause the results change slightly.

[3]http://www.sogou.com/labs/dl/tce.html

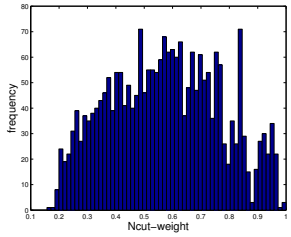

Figure 2: Distribution of Ncut-weights on Tweets data.

Table 2: *idf* and Ncut-weight behave different as in this example from the Twitter Tweets data

term	*idf*(rank)	Ncut-weight(rank)	Δrank
humidity	5.238(2054)	0.147(640)	+1414
pittsburgh	5.931(1454)	0.130(988)	+466
video	6.625(659)	0.200(161)	+498
cap	6.626(524)	0.141(764)	-240
org	6.114(1217)	0.108(1477)	-260
refuse	6.018(1380)	0.103(1578)	-198

4.2 Comparison with idf

As we saw in section 1, the *idf* weighting scheme has the problem of skew to high values in short texts. However, the Ncut-weight refrains from such problem by counting the term co-occurrence frequency instead of the document frequency. Figure 2 shows the distribution of Ncut-weights in Tweets data after preprocessed, which resembles a Gaussian distribution, and is much flatter than idf.

We further extracted some terms from the Tweets Data to explain that the two weighting schemes behave differently. In Table 2, the second and third columns show *idf* and the Ncut-weights of the example terms, respectively. Numbers in the parentheses denote the rank of terms while ordering by the weights in descending order. We can see that the first three words are very discriminative, but the weight is underestimated by *idf*. While the last three words with weak discriminability are overestimated by *idf*. Yet Ncut-weight scheme produces more reasonable weights of them.

4.3 Clusters Readability

We extracted four clusters from results on Tweets data as shown in Table 3. Top 5 weighted terms are presented in each cluster. It shows that: 1) some clusters found by Kmeans are not meaningful, like cluster 2 and cluster 3; 2) RLSI with *idf* is better than Kmeans, but still with some noise words in top weighted terms. For example, "february" ranks in the first place in cluster 3, but it is not directly related to weather; 3) "RLSI+nc" works better than "RLSI+idf"; 4) NMF with binary weights fails in cluster 4; 5) NMF with *idf* finds clusters more readable than "NMF+01". But some common words still rank higher than some others with more discriminability, like "social" and "medium" rank higher than "company"; 6) "NMF+nc" produces the most

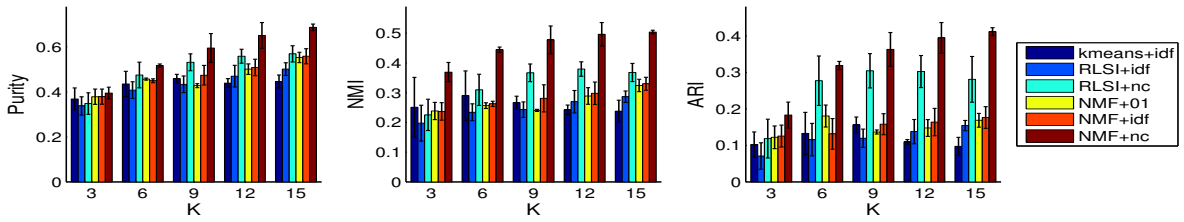

Figure 3: Comparison of (a)Purity, (b)NMI, (c)ARI w.r.t the cluster number k on Titles data

Table 3: Clusters generated by each methods on the Tweets data with $K = 15$

Methods	Kmeans+idf	RLSI+idf	RLSI+nc	NMF+01	NMF+idf	NMF+nc
cluster1: egyptian unrest†	egyptian	president	president	egypt	egyptian	egyptian
	egypt	egyptian	mubarak	egyptian	egypt	cairo
	mubarak	mubarak	egyptian	mubarak	mubarak	mubarak
	cairo	cairo	cairo	president	president	president
	protest	party	hosni	cairo	protester	protester
cluster2: market	player	market	market	market	market	market
	deal	sale	business	report	business	business
	market	plan	report	social	social	company
	review	business	medium	medium	medium	website
	press	party	social	online	company	social
cluster3: weather	super	february	temperature	wind	wind	temperature
	bowl	weather	humidity	humidity	humidity	humidity
	wind	temperature	barometer	rain	temperature	wind
	humidity	issue	hpa	temperature	rain	barometer
	temperature	humidity	mais	mph	mph	hpa
cluster4: football	green	buy	bowl	green	green	bowl
	bay	super	super	bay	bay	super
	packer	bowl	packer	packer	packer	packer
	super	party	bay	red	steelers	bay
	bowl	fan	xlv	yellow	pittsburgh	xlv

† cluster labels are assigned according to Top words in them manually

readable result. Besides, the results are very similar with "RLSI+nc" in these clusters.

4.4 Quantitative Evaluation

Since each document in the Titles data has a unique class label, we can evaluate the clustering results automatically. Three popular measures in clustering are used: purity, adjusted random index (ARI), and normalized mutual information (NMI)[2].

Figure 3 shows the clustering results of all the methods with respect to different K on Titles data. It is clear that "NMF+nc" outperforms all the baselines significantly in all evaluation metrics, especially when K is larger than 3. Besides, it is not surprise that "RLSI+nc" also shows great improvement than "RLSI+idf". Additionally, "NMF+idf" achieves slightly better result than "NMF+01", since the binary weights are least discriminative, while Kmeans works worst on these short texts.

5. CONCLUSION

Term weighting is important for NMF in document clustering. Conventional weighting schemes, like binary weights and *tfidf*, are not effective for short text clustering. We have proposed a novel term weight called Ncut-weight, which measures term's discriminability according to the words co-occurrences. The experiments show that the clustering performance of NMF is greatly improved with terms weighted by the Ncut-weight.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] R. Albright, J. Cox, D. Duling, A. Langville, and C. Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, NCSU Technical Report Math 81706. http://meyer. math. ncsu. edu/Meyer/Abstracts/Publications. html, 2006.

[2] C. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[3] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[4] Q. Wang, J. Xu, H. Li, and N. Craswell. Regularized latent semantic indexing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, pages 685–694. ACM, 2011.

[5] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[6] S. Yu and J. Shi. Multiclass spectral clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 313–319. IEEE, 2003.