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

Xiaohui Yan[1], Jiafeng Guo[1], Shenghua Liu[1], Xueqi Cheng[1], Yanfeng Wang[2]
[1]*Institute of Computing Technology, Chinese Academy of Sciences* , [2]*Sogou Inc.*
yanxiaohui@software.ict.ac.cn, {guojiafeng,liushenghua,cxq}@ict.ac.cn,wangyanfeng@sogou-inc.com

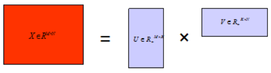
1. BACKGROUND

Documents Clustering by NMF

Non-negative matrix factorization (NMF) is a widely used document clustering method[Xu 2003], which decomopses the term-document matrix ***X*** into to low-rank non-negative matrices.

$$\min_{U\geq 0,V\geq 0} J(U,V)=\|X-UV\|^2$$

- X: each column represent a document via terms
- U: each column represent a topic via terms
- V: each coumen represent a document via topics



Term Weighting in NMF

- **Term weighting in term-doc matrix *X* is important for NMF**
  - Different representations of documents *X* will result in different factorized matrices *U* and *V*

- **tfidf is the most common term weighting scheme**

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

In a ducment, a term is more important/ discriminative
- if it occurs more often in the document
- if it occurs less in other documents

2. PROBLEM

Short Text

- Short texts are prevalent on the web
  - microblogs
  - SNS statuses
  - instant messages
  - …
- Short text clustering is important for various applications
  - emerging topics discovery
  - efficient index and retrieval personalized
  - recommendation
  - …

Problems of tfidf on Short Text

- tfidf always works well on normal text, but not on short text
  - most of terms usually occur only once in a short document
  - most of terms with a high idf value, due to the sparsity of data. Skewed distribution cannot discriminate terms very well.

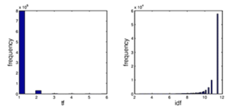


Figure 1: Frequency of (a) tf values, (b) idf values of terms in Tweets data set

3. OUR APPROACH

Ncut on Term Affinity Graph

- Consider a term affinity graph with adjacent matrix ***S*** = ***XX***^{T}, clustering terms is quivalents to cut graph G into K sub-graphs.
- A typical criterion to do that is called the normalized cut(Ncut) criterion, can be represented by the following trace maximization problem[Yu 2003]:

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

D is the diagonal degree matrix of S
U is an cluster indicator matrix,

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

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

Ncut-weighted NMF

- Theorem 1 suggest a new a term weighting matrix for matrix ***X***: ***D***$^{-1/2}$, i.e. the weight of term i is

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

A term is more important/ discriminative
- if it occurs less often in the corpus
- if it co-occurs less with other terms

- Ncut-weighted NMF

$$\min_{U\geq 0,V\geq 0} J(U,V)=\|Y-UV\|^2+\lambda(\|U\|^2+\|V\|^2)$$

4. EXPERIMENTS

Data Sets

- Data sets
  - Tweets, collected from twitter.com
  - Titles, news titles with assigned class labels from some news websites, which is published by Sogou Lab

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

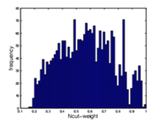


Figure 2:Distribution of Ncut-weights on Tweets data.

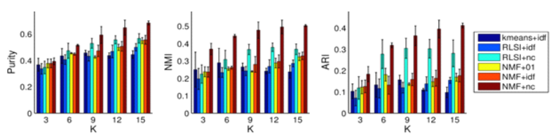
Comparison Ncut-weight with idf

- the Ncut-weight counts the term co-occurrence frequency instead of the document frequency.
  - Figure 2 shows Ncut-weights does not have the problem of skew to high values in short texts
  - Case study in Table 2 shows Ncut-weights captures terms' discriminative power better

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

5. CLUSTERING EVALUATION



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

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

6. CONCLUSIONS

Conclusions:

- Term weighting is important for NMF in document clustering. However, tranditional tfidf weights lost their discriminative power in short texts due to data sparsity.
- Nuct-weight, derived from Ncut algorithm on term affinity graph ,measures term's discriminability according to the words co-occurrence, avoiding the problem of tifidf on sparse term-document co-occurrence data.
- The experiments show that the clustering per-formance of NMF is greatly improved with terms weighted by the Ncut-weight.

References:

- 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.
- S. Yu and J. Shi. Multiclass spectral clustering. In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pages 313{319. IEEE, 2003.