Integrating Clustering and Multi-Document Summarization to Improve Document Understanding

Dingding Wang† Shenghuo Zhu‡ Tao Li† Yun Chi‡ Yihong Gong‡
† School of Computer Science
Florida International University
Miami, FL 33199
† {dwang003,taoli}@cs.fiu.edu

‡ NEC Laboratories. America, Inc.
10080 N. Wolfe Rd. SW3-350
Cupertino, CA 95014
‡{zsh,ychi,ygong}@sv.nec-labs.com

ABSTRACT

Document understanding techniques such as document clustering and multi-document summarization have been receiving much attention in recent years. Current document clustering methods usually represent documents as a term-document matrix and perform clustering algorithms on it. Although these clustering methods can group the documents satisfactorily, it is still hard for people to capture the meanings of the documents since there is no satisfactory interpretation for each document cluster. In this paper, we propose a new language model to simultaneously cluster and summarize the documents. By utilizing the mutual influence of the document clustering and summarization, our method makes (1) a better document clustering method with more meaningful interpretation and (2) a better document summarization method taking the document context information into consideration.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation, Performance

Keywords

Document Clustering, Multi-Document Summarization, Nonnegative Matrix Factorization with Given Bases

1. OUR METHOD

Document clustering and multi-document summarization are two fundamental tools for understanding documents. In this paper, we propose a new language model, factorization with given bases (FGB), by making use of both the term-document and term-sentence matrices to simultaneously cluster and summarize the documents. The proposed FGB model translates the clustering-summarization problem into minimizing the Kullback-Leibler divergence between the given documents and model reconstructed terms. The minimization process results two matrices which represent the probabilities of the documents and sentences given clusters (topics). The document clusters are generated by assigning each document to the topic with the highest probability, and the summary is formed with the sentences with the high probability in each topic.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

1.1 Framework Overview

Figure 1 shows the framework of our proposed document understanding system. First of all, the documents are preprocessed by removing formatting characters and stopping words. Then we apply the unigram language model to get the term-document matrix and the term-sentence matrix. Given the two matrices, our system performs nonnegative factorization on the term-document matrix using the term-sentence matrix as the bases. Upon convergence, the document-topic matrix and sentence-topic matrix are constructed, from which the document clusters and the corresponding summarized sentences can be generated simultaneously.

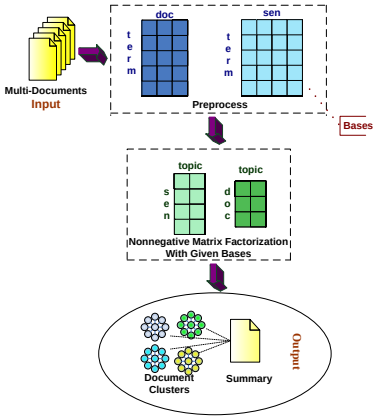

Figure 1: Overview of our proposed framework

1.2 FGB Model

The entire document set is denoted by $\mathcal{D}$. For each document $d \in \mathcal{D}$, we consider its language model,

$$p(w_1^n|\theta_d) = \prod_{i=1}^{n} p(w_i|\theta_d, w_1^{i-1}),$$

where θ_d denotes the model parameter for document d, w_1^n denotes the sequence of words $\{w_i \in \mathcal{W}\}_{i=1}^n$, i.e. the content of the document. $\mathcal{W}$ is the vocabulary. Similar to PLSI[4], we decompose the document language model into several common topic language models,

$$p(w_i|\theta_d, w_1^{i-1}) = \sum_{t\in\mathcal{T}} p(w_i|t, w_1^{i-1}) p(t|\theta_d, w_1^{i-1}),$$

where $\mathcal{T}$ is the set of topics. Here, we assume that given a topic, generating words is independent from the document, i.e.

$$p(w_i|t_i, \theta_d, w_1^{i-1}) = p(w_i|t_i, w_1^{i-1}).$$

Instead of freely choosing topic language models, we further assume that topic language models are mixtures of some existing *base language models*, i.e.

$$p(w_i|t, w_1^{i-1}) = \sum_{s\in\mathcal{S}} p(w|s, w_1^{i-1})p(s|t, w_1^{i-1}),$$

where $\mathcal{S}$ is the set of base language models. Here, we use sentence language models as the base language models. One benefit of this assumption is that each topic are represented by meaningful sentences, instead of directly by keywords.

For the sake of simplicity, we use unigram language models in this paper. Thus we have

$$p(w_i|\theta_d) = \sum_{t\in\mathcal{T}}\sum_{s\in\mathcal{S}} p(w_i|s)p(s|t)p(t|\theta_d).$$

We use the empirical distribution of observed sentences for $p(w|s)$, denoted by $\mathbf{B}_{w,s}$. The model parameters are $(\mathbf{U}, \mathbf{V})$, where

$$\mathbf{U}_{s,t} = p(s|t), \quad \mathbf{V}_{d,t} = p(t|\theta_d).$$

Thus, $p(w_i|\theta_d) = [\mathbf{BUV}^\top]_{w,d}$.

The FGB model uses the mixtures of some existing base language models as topic language models. When the base language models are language models with single word, then this model is identical to PLSI [4].

The parameter estimation is maximum likelihood estimation given occurrence of term-document. We use the empirical distribution of observed word-document distribution, $\tilde{p}(w, d)$, denoted by $\mathbf{A}_{w,d}$. The task is

$$\mathbf{U}, \mathbf{V} = \arg\min_{\mathbf{U},\mathbf{V}} \mathrm{KL}\left(\mathbf{A} \| \mathbf{BUV}^\top\right), \tag{1}$$

The algorithm for estimating FGB model is similar to NMF algorithms [5, 6, 3]. The computational algorithm of our model is shown in Algorithm 1.

Algorithm 1 Model factorization given base language models

Input: $\mathbf{A}$: term-document matrix.
$\mathbf{B}$: term-sentence matrix;
Output: $\mathbf{U}$: sentence-topic matrix;
$\mathbf{V}$: document-topic matrix.
begin
1. **Initialization:**
Randomly initialize $\mathbf{U}$ and $\mathbf{V}$ with normalizing each column of $\mathbf{U}$ and each row of $\mathbf{V}$ to 1
2. **Iteration:**
repeat
2.1 Compute $\mathbf{C}_{ij} = \mathbf{A}_{ij}/[\mathbf{BUV}^\top]_{ij}$;
2.2 Assign $\mathbf{U}_{st} \leftarrow \mathbf{U}_{st}\left[\mathbf{B}^\top\mathbf{CV}\right]_{st}$, and normalize each column to 1;
2.3 Compute $\mathbf{C}_{ij} = \mathbf{A}_{ij}/[\mathbf{BUV}^\top]_{ij}$;
2.4 Assign $\mathbf{V}_{dt} \leftarrow \mathbf{V}_{dt}\left[\mathbf{C}^\top\mathbf{BU}\right]_{dt}$, and normalize each row to 1;
until convergence
3. **Return U, V**
end

2. AN ILLUSTRATIVE EXAMPLE

In order to illustrate the clustering performance and interpretability of our FGB approach, in this example, we use the top 10 largest clusters of documents in TDT2 corpus to compare our clustering accuracy with the state-of-the-art methods and show the summarized description for each document cluster. Table 1 shows the clustering accuracy results and Table 2 shows the summaries for the top 4 largest document clusters.

k	2	4	6	8	10
K-means	0.6956	0.4081	0.3942	0.4024	0.3885
NMF	0.9848	0.7835	0.8120	0.8024	0.7547
FGB	**0.9869**	**0.8204**	**0.8535**	**0.8145**	**0.7811**
ITCC [2]	0.8766	0.6831	0.6652	0.6535	0.6480
MSRCC [1]	0.7346	0.4735	0.4730	0.4545	0.4464
ECC [1]	0.6838	0.4011	0.3972	0.3704	0.3613

Table 1: Accuracy comparison of different document clustering methods using TDT_10 data. Remark: k is the number of clusters; accuracy is calculated using the bestmap approach. "ITCC", "MSRCC", and "ECC" represent Information-theoretic, Minimum Sum-squared Residue, and Euclidean co-clustering algorithms, respectively.

topic1	- The Security Council has refused to lift the sanctions until Iraq complies with council resolutions demanding it destroy its weapons of mass destruction.
topic2	- Clinton says he had a very clear memory of the incident and he stands by the sworn court statement he has made that he did nothing wrong.
topic3	- The IOC had been expected to approve a new rule that all challenges to Olympic results must be make within three years after the games and settled by the time the next games begins.
topic4	- HONG KONG (AP): southeast Asian currencies hit new lows Tuesday for a second straight day, unnerving investors and sending regional stock markets tumbling.

Table 2: One-sentence summaries formed by our method for the top 4 largest topics in TDT2 corpus. Remarks: topic 1 corresponds to label 20015: Current Conflict with Iraq; topic 2 corresponds to label 20002: Monica Lewinsky Case; topic 3 corresponds to label 20013: 1998 Winter Olympics; and topic 4 corresponds to label 20001: Asian Economic Crisis.

Acknowledgements: The work of T. Li is partially supported by National Science Foundation under IIS-0546280, HRD-0317692, and IIP-0450552.

3. REFERENCES

[1] H. Cho, I. Dhillon, Y. Guan, and S. Sra. Minimum sum squared residue co-clustering. In *Proceedings of SIAM Data Mining 2004*.
[2] I. Dhillon, S. Mallela, and S. Modha. Information-theoretic co-clustering. In *Proceedings of SIGKDD 2001*.
[3] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of SIGKDD 2006*.
[4] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR 1999*.
[5] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS 2001*.
[6] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of SIGIR 2003*.