Integrating Document Clustering and Multidocument Summarization

DINGDING WANG, Florida International University
SHENGHUO ZHU, NEC Laboratories America
TAO LI, Florida International University
YUN CHI and YIHONG GONG, NEC Laboratories America

Document understanding techniques such as document clustering and multidocument summarization have been receiving much attention recently. Current document clustering methods usually represent the given collection of documents as a document-term matrix and then conduct the clustering process. Although many of these clustering methods can group the documents effectively, it is still hard for people to capture the meaning of the documents since there is no satisfactory interpretation for each document cluster. A straightforward solution is to first cluster the documents and then summarize each document cluster using summarization methods. However, most of the current summarization methods are solely based on the sentence-term matrix and ignore the context dependence of the sentences. As a result, the generated summaries lack guidance from the document clusters. In this article, we propose a new language model to simultaneously cluster and summarize documents by making use of both the document-term and sentence-term matrices. By utilizing the mutual influence of document clustering and summarization, our method makes; (1) a better document clustering method with more meaningful interpretation; and (2) an effective document summarization method with guidance from document clustering. Experimental results on various document datasets show the effectiveness of our proposed method and the high interpretability of the generated summaries.

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

Additional Key Words and Phrases: Document clustering, multidocument summarization, nonnegative matrix factorization with given bases

ACM Reference Format:
Wang, D., Zhu, S., Li, T., Chi, Y., and Gong, Y. 2011. Integrating document clustering and multidocument summarization. ACM Trans. Knowl. Discov. Data 5, 3, Article 14 (August 2011), 26 pages.
DOI = 10.1145/1993077.1993078 http://doi.acm.org/10.1145/1993077.1993078

1. INTRODUCTION

Document clustering and multidocument summarization are two fundamental tools for understanding document data and have attracted much attention in recent years.

The work is partially supported by a Florida International University (FIU) Dissertation Year Fellowship and NSF Grants IIS-0546280, CCF-0836359, and DMS-0915110.
Authors' addresses: D. Wang and T. Li, School of Computer Science, Florida International University, 11200 SW 8th St., Miami, FL 33199; email: {dwang003, taoli}@cs.fiu.edu; S. Zhu, Y. Chi, and Y. Gong, NEC Laboratories America, Inc., 10080 N. Wolfe Road, SW 3-350, Cupertino, CA 95014;
email: {zsh, ychi, ygong}@sv.nec-labs.com.

DOI 10.1145/1993077.1993078 http://doi.acm.org/10.1145/1993077.1993078

Given a collection of documents, document clustering aims to partition them into different groups called clusters; so that the documents in the same group are similar to each other, while the documents in different clusters are dissimilar [Duda et al. 2001]. Multidocument summarization, the process of generating a summary by reducing documents in size while retaining the main characteristics of the original documents, is another effective tool for understanding documents [Mani 2001; Ricardo and Berthier 1999]. Both of the document clustering and summarization techniques contribute to retrieving useful and meaningful information from documents, and they have a wide range of applications in information management and retrieval. For example, document clustering provides an efficient method for organizing and presenting Web search results [Zamir and Etzioni 1998], and the summarization used in snippet generation on the Web can assist users in further exploration [Turpin et al. 2007].

In order to perform document clustering, the document collection is usually represented as a document-term matrix where each row of the matrix represents a document and each column represents a term. Many clustering methods, such as K-means, hierarchical clustering algorithms and nonnegative matrix factorization (NMF) have been performed on the matrix to group the documents. However, these methods lack ability to interpret the documents because there is no satisfactory interpretation for each document cluster. Recently, coclustering algorithms were proposed to cluster documents and terms simultaneously using the dual relationship information between the documents and terms [Dhillon 2001; Dhillon et al. 2001]. In the coclustering framework, the document clusters are usually described and explained using representative terms in the associated term clusters. However, a list of representative words has limited interpretability due to the lack of semantic and context information. A more natural way to interpret each cluster is, to use the most important sentences extracted from the documents.

Multidocument summarization is an effective way to summarize each document cluster. In general, there are two types of summarization: extractive summarization and abstractive summarization [Jing and McKeown 2000; Knight and Marcu 2002]. Extractive summarization selects the important sentences from the original documents to form a summary, while abstractive summarization paraphrases the corpus using novel sentences. So, extractive summarization usually ranks the sentences in the documents according to their scores calculated by a set of predefined features, such as term frequency-inverse sentence frequency (TF-ISF) [Lin and Hovy 2001; Radev et al. 2004], sentence or term position [Lin and Hovy 2001; Yih et al. 2007], and number of keywords [Yih et al. 2007]. Abstractive summarization usually involves information fusion, sentence compression and reformulation [Jing and McKeown 2000; Knight and Marcu 2002]. Although an abstractive summary could be more concise, it requires deep natural language processing techniques. Thus extractive summaries are more feasible and practical. In this article, we study extractive summarization. Current document summarization methods usually represent the document collection as a sentence-term matrix where each row represents a sentence and each column represents a term. In some extractive summarization methods, clustering procedure is first used to generate sentence clusters and then a number of representative sentences are extracted from each sentence cluster. The problem with these methods is that they ignore the context dependency of the sentences and treat the sentences as independent of each other. In fact, the sentences within the same document or the same document cluster do have mutual influence, which should be utilized as additional knowledge to help the summarization.

In this article, we propose a new language model, factorization with given bases (FGB), where the given bases are sentence bases, by making use of both the

document-term and sentence-term matrices obtained from the original documents.[1] The model simultaneously clusters and summarizes the documents, and utilizes the mutual influence of the document clustering and summarization procedures to make, (1) a better document clustering method with more meaningful interpretation, and (2) an effective document summarization method taking the document context information into consideration. The FGB model translates the clustering-summarization problem into minimizing the Kullback-Leibler divergence between the given documents and the model reconstructed terms. The minimization process results in 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 by extracting the sentences with high probabilities in each topic.

The rest of the article is organized as follows. Section 2 discusses the related work of current methods in document clustering and multidocument summarization. Section 3 shows the overview of our framework. Our proposed FGB model and the computational algorithm are presented in Section 4. Comprehensive experimental results are shown in Section 5. Finally, Section 6 concludes the article.

2. RELATED WORK

2.1 Document Clustering

Document clustering has been widely studied. Traditional clustering techniques such as hierarchical and partitioning methods have been used in clustering documents. Hierarchical clustering proceeds successively by building a tree of clusters using bottom-up or top-down approaches. For example, hierarchical agglomerative clustering (HAC) [Duda et al. 2001] is a typical bottom-up hierarchical clustering method, which takes each document as a singleton cluster to start off with and then merges pairs of clusters until all clusters have been encapsulated into one final cluster that contains all documents. Partitioning methods attempt to directly decompose the document collection into a number of disjoint classes such that the documents in a cluster are more similar to one another than the documents in other clusters [He et al. 2004; Liu et al. 2003]. For example, K-means [Duda et al. 2001] is a typical partitioning method, which aims to minimize the sum of the squared distances between the documents and the corresponding cluster centers.

Model-based clustering techniques have also been used in document clustering [Elkan 2006; Zhong and Ghosh 2003, 2005], where clusters are represented as probabilistic models in a model space that is conceptually separated from the data space. Probabilistic Latent Semantic Indexing (PLSI) is an unsupervised learning method based on statistical latent class models and has also been successfully applied to document clustering [Hoffman 1999]. PLSI is further developed into a more comprehensive Latent Dirichlet Allocation model [Blei et al. 2002]. Recently, graph-based clustering algorithms have emerged as a promising clustering approach [Wang et al. 2007]. The basic idea of graph-based clustering is to model the dataset as a weighted graph in which each data point is represented as a node and the weights on the edges represent the similarities between the corresponding nodes. Then the clustering is achieved by optimizing some predefined criterion functions on the graph. For instance, spectral clustering is one of the most representative graph-based clustering

[1]A preliminary report of this work was published as a 2-page poster in Wang et al. [2008b]. The journal manuscript provides detailed description, in-depth theoretical analysis and comprehensive experimental results.

approaches, which aims to optimize cut criteria such as normalized cut [Shi and Malik 2000; Zha et al. 2001], ratio cut [Zien et al. 1999], and min-max cut [Ding et al. 2001], and so on. These criteria can usually be optimized via eigen-decomposition. One of the major problems of the spectral clustering approach is that the eigenvectors obtained by the decomposition have negative values, which are hard to interpret. The nonnegative matrix factorization (NMF) methods are proposed [Ding et al. 2006; Lee and Seung 2001; Li and Ding 2006; Xu and Gong 2004; Xu et al. 2003], which factorize the document-term matrix and constrain each value to be nonnegative. NMF methods produce better semantic relationships between the decomposition results and the structure of the data.

Although all of these clustering techniques can effectively group the documents into different clusters, there is no natural interpretation for each document cluster, which may create a barrier for people to understand the documents clearly.

Coclustering algorithms are proposed, which aim at clustering document and term simultaneously by making use of the dual relationship information [Dhillon 2001; Dhillon et al. 2001; Li 2005]. Bipartite spectral graph partitioning approaches are proposed in Zha et al. [2001] to cocluster words and documents. Long et al. [2006] propose a general principled model, called Relation Summary Network, to cocluster the heterogeneous data on a k-partite graph. Subspace clustering algorithms have also been developed for discovering low-dimensional clusters in high-dimension document space [Jing et al. 2007; Li et al. 2004]. Although in these algorithms, document clusters can be described using representative terms, the interpretability of the terms is limited and not natural.

2.2 Document Summarization

Multidocument summarization is used to extract the main ideas of the documents and put them into a short summary. As mentioned in Section 1, here we only consider extractive multidocument summarization. Extractive summarization can be either generic or query-relevant. *Generic document summarization* should reflect the major content of the documents without any additional information. *Query-relevant document summarization* should focus on the information expressed in the given queries, i.e., the summaries must be biased to the given queries. Both generic and query-relevant document summarization have been studied recently. Gong and Liu [2001] propose a method using latent semantic analysis (LSA) to select sentences with high ranking for summarization. Goldstein et al. [1999] propose a maximal marginal relevance (MMR) method to summarize documents based on the cosine similarity between the query and the sentences. Other approaches include NMF-based summarization [Park et al. 2007], CRF-based summarization [Shen et al. 2007], and a hidden Markov model (HMM) based method [Conroy and O'Leary 2001]. In addition, graph-ranking based methods are also proposed. For example, an algorithm called LexPageRank is proposed in Erkan and Radev [2004] to compute the sentence importance based on the concept of eigenvector centrality (prestige). Other graph-based summarization methods have been proposed in Mihalcea and Tarau [2005] and Wan and Yang [2008].

There are also some newly developed summarization systems focusing on query-relevant document summarization. Tang et al. [2009] incorporate the query information into a topic model, and use the topic-based scores and term frequency to estimate the importance of the sentences. Wang et al. [2008a] calculate sentence-sentence similarities by sentence-level semantic analysis, and cluster the sentences via symmetric nonnegative matrix factorization. Wei et al. [2008] extend the mutual reinforcement principle between sentences and terms to the document-sentence-term mutual reinforcement chain, and use query-sensitive similarity to measure the

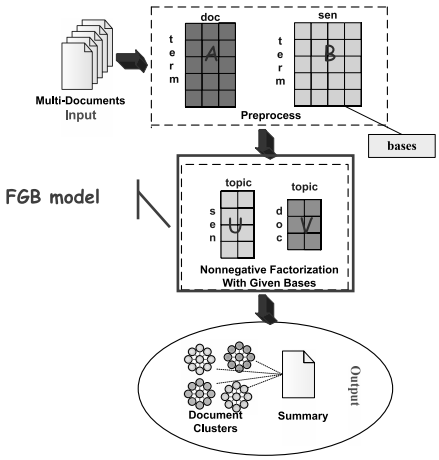

Fig. 1. An overview of our proposed framework.

affinity between the texts. Wan and Xiao [2009] use a manifold-ranking algorithm by considering the within-document sentence relationships and the cross-document sentence relationships as two separate modalities (graphs). Nastase [2008] uses Wikipedia as external knowledge to expand queries and build the connection between the queries and the sentences in documents. However, all these existing approaches conduct sentence clustering and extraction based on the sentence-term matrix only. A very recent short paper [Wang et al. 2009] has shown that information on the document side can benefit the summarization performance. It would be necessary to discuss the relation between document space and sentence space in detail.

There are also a few papers exploring the relationship between clustering and summarization [Dunlavy et al. 2007; Mana-Lopez et al. 2004; McKeown et al. 2002]. Mana-Lopez et al. [2004] conduct multidocument summarization based on the document clustering results, and generate two kinds of summaries covering the common parts of the documents and the particularities of each document respectively. This article aims to compare document clusters and find the commonality between them. Dunlavy et al. [2007] implement a hybrid information retrieval system to perform queries, clustering, and summarization, where the information derived from the query determines the document cluster assignment and the summarization is based on a Hidden Markov Model (HMM) as proposed in Conroy and O'Leary [2001]. McKeown et al. [2002] developed a system for news event detection and summarization, which follows a pipeline architecture to first cluster articles into event clusters, and then put them into different categories, and for each category one type of summarization strategies is performed. However, in these systems, the clustering and summarization are two consecutive steps, while in our work we simultaneously obtain the document clusters and the summaries so the mutual influence of the two procedures can be well utilized.

3. THE FRAMEWORK OVERVIEW

Figure 1 shows the framework of our proposed document-understanding system, which aims to help people easily understand documents by simultaneously clustering and summarizing multiple documents. First of all, the documents are preprocessed by removing formatting characters and stop words. Then we apply the unigram language

model to get the document-term matrix and the sentence-term matrix. Given the two matrices, our system performs nonnegative factorization on the document-term matrix using the sentence-term matrix as the basis. Upon convergence, the document-topic matrix and sentence-topic matrix are obtained, from which the document clusters and the corresponding summaries can be generated simultaneously. Section 4 will describe the FGB model, which is the core module of our system, in detail.

4. THE 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 and w_1^n denotes the sequence of words $\{w_i \in \mathcal{W}\}_{i=1}^n$—the content of the document. $\mathcal{W}$ is the vocabulary. Similar to PLSI [Hoffman 1999], 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:

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

Instead of freely choosing topic language models, we further assume that topic language models are mixtures of some existing base language models:

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

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 is represented by meaningful sentences, instead of directly by keywords.

For a trade-off between simplicity and accuracy, we use unigram language models in this article to model the process of generating sentences from the point view of summarization. In addition, unigram language models are most commonly used in Information Retrieval (IR). Usually, the performance of the retrieval does not directly depend on the structure of sentences, so unigram language models are sufficient for most of the IR tasks instead of complex language models. 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 $p(w|s)$, and let $\mathbf{B}_{w,s} = \tilde{p}(w|s)$. 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, $p(w_i|\theta_d) = [\mathbf{BUV}^\top]_{w,d}$.

Now we consider the prior distribution of $\mathbf{U}$ and $\mathbf{V}$. We take the Dirichlet distribution, the conjugate prior of the multinomial distribution,

$$\mathbf{U}_{\cdot,t} \sim \mathrm{Dir}(\alpha^U_{\cdot,t} + 1), \quad \mathbf{V}_{d,\cdot} \sim \mathrm{Dir}(\alpha^V_{d,\cdot} + 1).$$

We assume that the total number of pseudo instances is α, and they are evenly distributed on all possible tuples. Since $\alpha^U_{s,t}$ and $\alpha^V_{d,t}$ are considered as the number of pseudo-tuples, we have

$$\alpha^U_{s,t} = \alpha/|\mathcal{S}|, \quad \alpha^V_{d,t} = \alpha/|\mathcal{T}|.$$

The parameter estimation is an MAP estimation (or maximum likelihood estimation) of N observed documents, and $\mathbf{A}_{w,d} = N\tilde{p}(w,d)$. The task is

$$\mathbf{U}, \mathbf{V} = \arg\min_{\mathbf{U},\mathbf{V}} \ell(\mathbf{U}, \mathbf{V}), \tag{1}$$

where $\ell(\mathbf{U}, \mathbf{V}) = \mathrm{KL}\left(\mathbf{A} \| \mathbf{B}\mathbf{U}\mathbf{V}^\top\right) - \ln \Pr(\mathbf{U}, \mathbf{V})$. Here, KL divergence is used to measure the difference between the distributions of $\mathbf{A}$ and the estimated $\mathbf{B}\mathbf{U}\mathbf{V}^\top$.

4.1 Computational Algorithm

Similar to the nonnegative matrix factorization algorithm in Lee and Seung [2001], we have the following Theorem.

THEOREM 1. *The loss $\ell(\mathbf{U}, \mathbf{V})$ as Eq. (1) is nonincreasing under the update rules,*

$$\mathbf{U}_{st} \leftarrow \beta_t \left\{ \mathbf{U}_{st} \left[\mathbf{B}^\top \mathbf{C} \mathbf{V} \right]_{st} + \alpha_U \right\},$$
$$\mathbf{V}_{dt} \leftarrow \gamma_t \left\{ \mathbf{V}_{dt} \left[\mathbf{C}^\top \mathbf{B} \mathbf{U} \right]_{dt} + \alpha_V \right\},$$

where $\mathbf{C}_{ij} = \mathbf{A}_{ij}/[\mathbf{B}\mathbf{U}\mathbf{V}^\top]_{ij}$, β_t's are normalizing parameters to make $\sum_s \mathbf{U}_{st} = 1$ and γ_d's are normalizing parameters to make $\sum_t \mathbf{V}_{dt} = 1$.

PROOF. Let $\alpha_{iklj} = \mathbf{B}_{ik}\widetilde{\mathbf{U}}_{kl}\widetilde{\mathbf{V}}_{jl}/[\mathbf{B}\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top]_{ij}$. Applying Jensen's inequality, we obtain

$$\begin{aligned}
\ell(\mathbf{U}, \mathbf{V}) &= \sum_{ij}\Big(\sum_{kl} \mathbf{B}_{ik}\mathbf{U}_{kl}\mathbf{V}_{jl} - \mathbf{A}_{ij} \ln\Big(\sum_{kl} \mathbf{B}_{ik}\mathbf{U}_{kl}\mathbf{V}_{jl}\Big)\Big) \\
&\quad -\alpha_U \sum_{kl} \ln \mathbf{U}_{kl} - \alpha_V \sum_{jl} \ln \mathbf{V}_{jl} + c_1 \\
&\le \sum_{ij}\sum_{kl}\Big(\mathbf{B}_{ik}\mathbf{U}_{kl}\mathbf{V}_{jl} - \alpha_{iklj}\mathbf{A}_{ij} \ln \frac{\mathbf{B}_{ik}\mathbf{U}_{kl}\mathbf{V}_{jl}}{\alpha_{iklj}}\Big) \\
&\quad -\alpha_U \sum_{kl} \ln \mathbf{U}_{kl} - \alpha_V \sum_{jl} \ln \mathbf{V}_{jl} + c_1 \\
&= -\sum_{ijkl} \mathbf{C}_{ij}\mathbf{B}_{ik}\widetilde{\mathbf{U}}_{kl}\widetilde{\mathbf{V}}_{jl} \ln(\mathbf{U}_{kl}\mathbf{V}_{jl}) - \alpha_U \sum_{kl} \ln \mathbf{U}_{kl} \\
&\quad -\alpha_V \sum_{jl} \ln \mathbf{V}_{jl} + c_2 \\
&\stackrel{\mathrm{def}}{=} \mathcal{Q}(\mathbf{U}, \mathbf{V}; \widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}).
\end{aligned} \tag{2}$$

Algorithm 1 Model factorization given base language models

Input: **A**: term-document matrix.
B: term-sentence matrix;
Output: **U**: sentence-topic matrix;
V: document-topic matrix.
begin
1. **Initialization:**
Initialize **U** and **V** to follow Dirichlet distribution,
with hyper-parameter α_U and α_V, respectively.
2. **Iteration:**
repeat
2.1 Compute $\mathbf{C}_{ij} = \mathbf{A}_{ij}/[\mathbf{BUV}^\top]_{ij}$;
2.2 Assign $\mathbf{U}_{st} \leftarrow \mathbf{U}_{st}\left[\mathbf{B}^\top\mathbf{CV}\right]_{st} + \alpha_U$,
and normalize each column to 1;
2.3 Compute $\mathbf{C}_{ij} = \mathbf{A}_{ij}/[\mathbf{BUV}^\top]_{ij}$;
2.4 Assign $\mathbf{V}_{dt} \leftarrow \mathbf{V}_{dt}\left[\mathbf{C}^\top\mathbf{BU}\right]_{dt} + \alpha_V$,
and normalize each row to 1;
until convergence
3. **Return U, V**
end

The equality holds when $\mathbf{U} = \widetilde{\mathbf{U}}$ and $\mathbf{V} = \widetilde{\mathbf{V}}$. Instead of minimizing ℓ, we minimize $\mathcal{Q}$ without the nonnegative constraints. Later on, we find that the update rules satisfy the nonnegative constraints. The Lagrangian of $\mathcal{Q}$ is

$$\mathcal{L}(\mathbf{U}, \mathbf{V}; \boldsymbol{\xi}) = \mathcal{Q}(\mathbf{U}, \mathbf{V}; \widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}) + \boldsymbol{\xi}^\top(\mathbf{U}^\top\mathbf{1} - \mathbf{1}). + \boldsymbol{\zeta}^\top(\mathbf{V1} - \mathbf{1}). \tag{3}$$

The Karush-Kuhn-Tucker (KKT) conditions are:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{U}_{kl}} = -\frac{1}{\mathbf{U}_{kl}}\widetilde{\mathbf{U}}_{kl}\left[\mathbf{B}^\top\mathbf{C}\widetilde{\mathbf{V}}\right]_{kl} - \frac{\alpha_U}{\mathbf{U}_{kl}} + \xi_l = 0, \tag{4}$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{V}_{jl}} = -\frac{1}{\mathbf{V}_{jl}}\widetilde{\mathbf{V}}_{jl}\left[\mathbf{C}^\top\mathbf{B}\widetilde{\mathbf{U}}\right]_{jl} - \frac{\alpha_V}{\mathbf{V}_{jl}} + \zeta_j = 0, \tag{5}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\xi}_l} = \sum_k \mathbf{U}_{kl} - 1 = 0, \tag{6}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\zeta}_j} = \sum_l \mathbf{V}_{jl} - 1 = 0. \tag{7}$$

We derive the update rule from the KKT conditions. We can verify that the update rules keep **U** and **V** nonnegative. □

Based on Theorem 1, the computational algorithm of our model is Algorithm 1.

4.2 Relation with Other Models

The FGB model uses mixtures of some existing base language models as topic language models. When the base language models are language models with single words, then this model is identical to PLSI [Hoffman 1999] (or LDA [Blei et al. 2002]). In this case,

Table I. A Simple Example with Eight Sentences. D_i Represents the i^{th} Document and S_j is the j^{th} Sentence

D_1	S_1:	Apple is a corporation manufacturing consumer electronics.
	S_2:	Apple's design seems a lot more revolutionary to most Americans.
D_2	S_3:	Apple is a corporation manufacturing consumer electronics.
	S_4:	The design of Apple's products is more revolutionary than others in the market.
D_3	S_5:	Apple is a corporation manufacturing consumer electronics.
	S_6:	The prices of Apple's are relatively high.
D_4	S_7:	Apple is a corporation manufacturing consumer electronics.
	S_8:	With the performance, Apple's machines have higher price.

the matrix $\mathbf{B}$ is the identity matrix, and the algorithm is the same as NMF with KL divergence loss [Lee and Seung 2001] when the α's are zeros.

Inspired by the NMF algorithm, which uses the Frobenius norm in Lee and Seung [2001], we can also replace the KL divergence loss with the Frobenius norm, since KL divergence is derived from a probabilistic generative model, which has an explicit explanation for each parameter:

$$\ell_F(\mathbf{U}, \mathbf{V}) = \|\mathbf{A} - \mathbf{B}\mathbf{U}\mathbf{V}^\top\|_F^2.$$

When the matrix $\mathbf{B}$ is the identity matrix, this loss derives the NMF algorithm with the Frobenius norm in Lee and Seung [2001]. This formula was used for document clustering in Xu et al. [2003] and Ding et al. [2006]. In Xu and Gong [2004], $\mathbf{B} = \mathbf{A}$ is used to derive another clustering loss, which is analogous to using documents themselves as the base language models.

4.3 An Illustrative Example

To demonstrate the advantages of our proposed FGB model, a simple example is given in Table I. The synthetic dataset contains four very short articles, each of which has only two sentences (8 sentences in total). We cluster the documents into two groups and also provide a one-sentence summary for each document cluster.

Looking at the data directly, we know that D_1 and D_2 present the nice design of Apple's products, and D_3 and D_4 are related to the high prices. It is easy to group D_1 and D_2 into one cluster and D_3 and D_4 into the other using any existing clustering algorithm. However, only our proposed FGB model can generate a sentence-level interpretation of each document cluster. Now we apply our computational algorithm to this example. After removing stop words, we use 15 terms to construct the document-term and term-sentence matrices. They are "Apple," "corporation," "manufacturing," "consumer," "electronics," "design," "revolutionary," "Americans," "products," "market," "others," "performance," "high," "price," and "machines" (denoted as $w_1 \sim w_{15}$ respectively).

The input matrices are:

$$\mathbf{A}^\top = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ D_4 \end{array} \begin{array}{c} \begin{array}{cccccccccccccc} w_1 & & & & & & & \cdots & & & & & & w_{15} \end{array} \\ \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array},$$

where each row of A is a term (as listed), each column corresponds to a document, and each element is the number of times that the corresponding term appears in the corresponding document; and

$$\mathbf{B} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \\ w_{10} \\ w_{11} \\ w_{12} \\ w_{13} \\ w_{14} \\ w_{15} \end{array} \begin{array}{c} \begin{array}{cccccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 & S_8 \end{array} \\ \begin{pmatrix} 0.20 & 0.25 & 0.20 & 0.17 & 0.20 & 0.33 & 0.20 & 0.20 \\ 0.20 & & 0.20 & & 0.20 & & 0.20 & \\ 0.20 & & 0.20 & & 0.20 & & 0.20 & \\ 0.20 & & 0.20 & & 0.20 & & 0.20 & \\ 0.20 & & 0.20 & & 0.20 & & 0.20 & \\ & 0.25 & & 0.17 & & & & \\ & 0.25 & & 0.17 & & & & \\ & 0.25 & & & & & & \\ & & & 0.17 & & & & \\ & & & 0.17 & & & & \\ & & & 0.17 & & & & \\ & & & & & & & 0.20 \\ & & & & & 0.33 & & 0.20 \\ & & & & & 0.33 & & 0.20 \\ & & & & & & & 0.20 \end{pmatrix} \end{array},$$

where each row of B represents a term and each column represents a sentence. Note that each column of $\mathbf{B}$ has been normalized to 1.

We randomly initialize U and V and normalize them accordingly. After convergence and normalization, we obtain:

$$\mathbf{U} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8 \end{array} \begin{array}{c} \begin{array}{cc} \text{topic 1} & \text{topic 2} \end{array} \\ \begin{pmatrix} 0.08 & 0.09 \\ 0.00 & 0.13 \\ 0.22 & 0.11 \\ 0.00 & 0.38 \\ 0.21 & 0.20 \\ 0.00 & 0.00 \\ 0.06 & 0.09 \\ 0.43 & 0.00 \end{pmatrix} \end{array} \quad \text{and} \quad \mathbf{V} = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ D_4 \end{array} \begin{array}{c} \begin{array}{cc} \text{topic 1} & \text{topic 2} \end{array} \\ \begin{pmatrix} 0.00 & 1.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \end{pmatrix} \end{array}.$$

From V, we observe that D_1 and D_2 belong to one cluster and D_3 and D_4 belong to the other. And from U, S_4 and S_8 are two sentences with the highest probability in each topic, so they are selected to generate the summary for each document cluster respectively. Note that although sentences 1, 3, 5, and 7 are the same, since we randomly initialize U, the exact values in the final U for these sentences are not equal. However the selection of the most relevant sentence for each topic is stable after the convergence of our algorithm. The generated summaries are listed in Table III, and we see that one cluster is related to the design of Apple's products, and the other is related to the high price, which are consistent with human observation and perception.

One may argue that we can easily use a two-step approach by clustering the documents first and then using current document summarization methods to summarize each document cluster. However, this solution makes the two procedures independent, and we believe the mutual influence between document clustering and sentence selection will help each other. Table II shows the summaries formed by the two-step approach where the most centrally located sentence in each cluster using cosine similarity (e.g., S_1 and S_5 for the two clusters) is chosen as the summary sentence. (Note that here "centrally locate" sentences means the sentences most similar to the rest,

Table II. Summaries Formed by the Two-Step Approach

D_1&D_2	Apple is a corporation manufacturing consumer electronics.
D_3&D_4	Apple is a corporation manufacturing consumer electronics.

Table III. Summaries Formed by Our Proposed Algorithm Procedure

D_1&D_2	The design of Apple's products is more revolutionary than others in the market.
D_3&D_4	With the similar performance, Apple's machines have higher price than others.

Table IV. Representative Terms Selected by the NMF Algorithm

D_1&D_2	Apple, corporation, manufacturing
D_3&D_4	Apple, performance, higher

not the physical location.) Sentences S_1 and S_5, selected by the two-step approach are not informative since they do not contain information on either the price or the design. If we use the NMF algorithm on the document-term matrix to select representative terms to interpret each cluster, we obtain the results listed in Table IV. Comparing the results in Table III with our summaries in Table II, we can see that the interpretability of terms is limited and unsatisfactory. It is obvious that sentences express more information than the words, and also make a more natural description on the documents. Another observation is that most of the keywords picked by NMF in Table IV appear in the selected sentences using our proposed method in Table III, while the sentences make the summaries much more natural than the selected words.

5. EXPERIMENTS

In this section, we conduct two sets of experiments to evaluate our document clustering results and the quality of the simultaneously generated summaries. Section 5.1 examines the effectiveness of our system on document clustering, and Section 5.2 evaluates the quality of the generated summaries.

5.1 Experiments on Document Clustering

5.1.1 Dataset. In order to compare our document clustering method with the state-of-the-art methods, we use TDT2 and Reuters-21578 document corpora. They are two of the most ideal document collections for document clustering since the documents in them have been manually labeled. The TDT2 corpora contains 100 clusters of 64527 documents from six news agencies including American Broadcasting Company (ABC), Cable News Network (CNN), Voice of America (VOA), New York Times (NYT), Public Radio International (PRI), and APW. In our experiments, we use TDT_10, a subset of TDT2, which contains the top 10 largest clusters of documents. The Reuters corpora contains 21578 documents which are grouped into 135 clusters, and we use the top 10 largest clusters to get the Reuters_10 dataset. Table V describes the statistics of these datasets. Note that there is a larger dataset of the Reuters corpora called RCV-1. However, we use Reuters-21578 instead of the RCV-1 because the RCV-1 data are multilabeled and of hierarchy structure. Though we performed some classification tasks on RCV-1, our understanding with the label structure of RCV-1 is not as good as with of Reuters-21578.

It should be pointed out that our algorithm is not limited to small number of clusters. In fact, we have run our algorithm to cluster very large scale datasets, such as blog data with 2 million documents and more than 27 million sentences. We did not include the clustering results on blog data because we do not have the true labels of

Table V. Description of the Datasets Used for Document Clustering

	TDT_10	Reuters_10
Number of Documents Used	7879	7754
Number of Clusters	10	10

those documents and thus it is difficult to evaluate the results. Hence in our clustering experiments, we only report experimental results on the benchmark datasets with ground truth.

5.1.2 Baselines. To compare our clustering results with the state-of-art methods, we implement the following document clustering methods as the baselines. K-means and NMF are popular methods used in document clustering. And since coclustering algorithms can cluster documents and terms simultaneously by making use of the dual relationship information, we also compare our method with three widely used coclustering algorithms. The algorithms used in the experiment comparison are listed in the following.

— *KM.* the traditional K-means Algorithm [Duda et al. 2001].
— *NMF.* document clustering based on nonnegative matrix factorization [Xu et al. 2003].
— *ITCC.* the information-theoretic coclustering algorithm [Dhillon et al. 2001].
— *MSRCC.* the minimum sum-squared residue coclustering algorithm [Cho et al. 2004].
— *ECC.* the Euclidean coclustering algorithm [Cho et al. 2004].

5.1.3 Evaluation Methods. To measure the clustering quality, we use accuracy and normalized mutual information (NMI) as performance measures.

— Accuracy measures the relationship between each cluster and the ground truth class. It sums up the total matching degree between all pairs of clusters and classes. Accuracy can be represented as:

$$Accuracy = Max\left(\sum_{C_k, L_m} T(C_k, L_m)\right)/N,$$

where C_k denotes the k-th cluster, and L_m is the m-th class. $T(C_k, L_m)$ is the number of entities that belong to class m and are assigned to cluster k. Accuracy computes the maximum sum of $T(C_k, L_m)$ for all pairs of clusters and classes, and there is no overlap among these pairs. It is obvious that the greater the accuracy, the better the clustering performance.

— NMI [Strehl and Ghosh 2003] measures the amount of statistical information shared by two random variables representing cluster assignment and underlying class label. Suppose entry n_{ij} denotes the amount of data items belonging to cluster i and class j. NMI is then computed as:

$$NMI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{k}\frac{n_{ij}}{n}log\frac{n_{ij}n}{n_{i.}n_{.j}}}{\sqrt{(\sum_{i=1}^{c}\frac{-n_{i.}}{n}log\frac{n_{i.}}{n})(\sum_{j=1}^{k}\frac{-n_{.j}}{n}log\frac{n_{.j}}{n})}},$$

where $n_{i.} = \sum_{j=1}^{k} n_{ij}$, $n_{.j} = \sum_{i=1}^{c} n_{ij}$, n, c, k denote the total number of data objects, the number of clusters, and the number of classes, respectively. Based on our prior knowledge of the number of classes, we set the number of clusters equal to the true number of classes: $c = k$.

Table VI. Accuracy Comparison of Different Document Clustering Methods Using TDT_10. Remark: k is the Number of Clusters

k	2	4	6	8	10
KM	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	0.8766	0.6831	0.6652	0.6535	0.6480
MSRCC	0.7346	0.4735	0.4730	0.4545	0.4464
ECC	0.6838	0.4011	0.3972	0.3704	0.3613

Table VII. NMI Comparison of Different Document Clustering Methods Using TDT_10

k	2	4	6	8	10
KM	0.1175	0.1702	0.2521	0.2913	0.3004
NMF	0.8828	0.6928	0.7354	0.7266	0.6934
FGB	**0.8959**	**0.6986**	**0.7536**	**0.7342**	**0.7077**
ITCC	0.7325	0.6228	0.6471	0.6682	0.6685
MSRCC	0.2735	0.2956	0.3144	0.3276	0.3227
ECC	0.3058	0.2726	0.3426	0.3572	0.3464

Table VIII. Accuracy Comparison of Different Document Clustering Methods Using Reuters_10

k	2	4	6	8	10
KM	0.5258	0.4915	0.4243	0.3808	0.3745
NMF	0.6836	0.6788	0.4422	0.4233	0.4533
FGB	**0.6883**	**0.6829**	**0.4848**	**0.4736**	**0.4746**
ITCC	0.5538	0.5158	0.4443	0.3963	0.3833
MSRCC	0.4606	0.4025	0.3482	0.3308	0.3366
ECC	0.3883	0.3268	0.3078	0.2833	0.2885

5.1.4 Experimental Results. In this set of experiments, we compare our methods with the five baseline document clustering methods as described in Section 5.1.2.

5.1.4.1 Overall Clustering Performance. First of all, we evaluate the overall accuracy and NMI performance of different clustering methods. We vary the cluster number k from 2 to 10 by selecting the top k largest clusters from the two datasets, respectively. Tables VI–IX show the accuracy and NMI results using different methods on TDT_10 and Reuters_10 datasets. We use Figures 2–5 to illustrate the results visually.

Our proposed method outperforms the K-means algorithm because of the flexibility of matrix factorization, which can model widely varying data distributions. The objective function of K-means clustering only attempts to capture rigid spherical clusters. When the data distribution is far from spherical, the matrix factorization has obvious advantages. Our method utilizes the sentence-term matrix as the given bases to conduct matrix factorization.

Note that our proposed factorization model is fundamentally different from the factorization model used in Xu et al. [2003], where the the input document-term matrix X is factorized into two nonnegative matrices F and G, where F gives document clustering information and G provides cluster centroid information. In their follow-up work,

Table IX. NMI Comparison of Different Document Clustering Methods Using Reuters_10

k	2	4	6	8	10
KM	0.0206	0.0480	0.0732	0.0958	0.1161
NMF	0.2404	0.2391	0.2229	0.2384	0.2451
FGB	**0.2447**	**0.2436**	**0.2706**	**0.2789**	**0.2708**
ITCC	0.1836	0.2028	0.1622	0.1818	0.1839
MSRCC	0.0404	0.0560	0.0786	0.0924	0.1248
ECC	0.0264	0.0436	0.0686	0.0744	0.0985

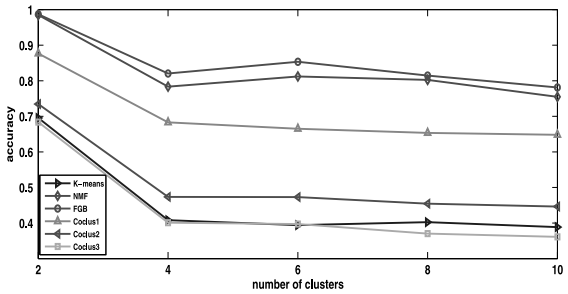

Fig. 2. Accuracy results of different clustering methods using TDT_10.

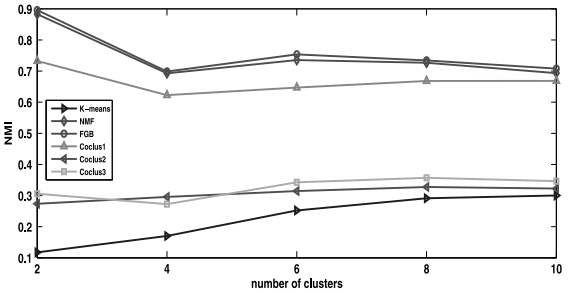

Fig. 3. NMI results of different clustering methods using TDT_10.

Xu and Gong proposed that F can be represented as a linear transformation of X, i.e., $F = XW$, and hence derived a 3-factor factorization [Xu and Gong 2004]. In our factorization model, we have an extra factor to absorb the different scales of X, F, G and provide additional degrees of freedom such that the low-rank matrix representation of X remains accurate while F gives document clustering information and G gives sentence-term information. Our proposed factorization offers a natural framework for simultaneous document clustering and sentence-level document summarization. In addition, the factorization model in Xu et al. [2003] and Xu and Gong [2004] uses the Frobenius norm, which has no clear generative process from the probabilistic point of view. In contrast, we derive KL divergence from a probabilistic generative model, which has an explicit explanation for each parameter. Moreover, in the factorization model of Xu et al. [2003] and Xu and Gong [2004], the information of sentences is not used, thus the results of NMF cannot directly generate the summarization. Our method is able to interpret the document clusters by the simultaneously generated summaries. The case study in Section 5.1.4.3 examines the interpretability of our method.

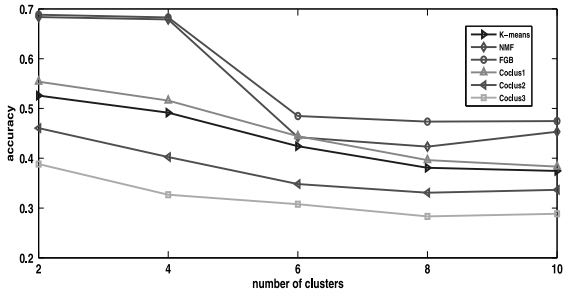

Fig. 4. Accuracy results of different clustering methods using Reuters_10.

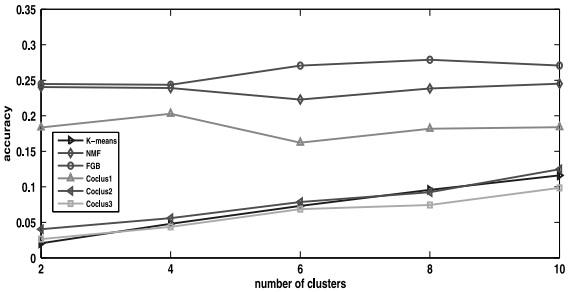

Fig. 5. NMI results of different clustering methods using Reuters_10.

Table X. The Statistical Significance Matrix Using Accuracy. Remark: "$\gg$" ("$\ll$") Indicates that Methods of the Row Perform Significantly Better (Worse) than the Methods of the Column, and "$>$" ("$<$") Indicates the Relationship is Not Significant. For All Statistical Significance Tests, the Minimum Confidence Interval is 95%

	KM	NMF	ITCC	MSRCC	ECC	FGB
KM	-	$\ll$	$\ll$	$>$	$>$	$\ll$
NMF	$\gg$	-	$>$	$\gg$	$\gg$	$\ll$
ITCC	$\gg$	$<$	-	$\gg$	$\gg$	$\ll$
MSRCC	$<$	$\ll$	$\ll$	-	$\gg$	$\ll$
ECC	$<$	$\ll$	$\ll$	$\ll$	-	$\ll$
FGB	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	-

5.1.4.2 Statistical Significance. In order to statistically compare the performance of our FGB model with other clustering methods, we use paired t-test based statistical tests to determine the significance of our results [Devore and Peck 1977]. We sample each dataset (for each k) ten times, and 70% of documents from the original dataset are randomly selected as a sample dataset. All the comparisons are based on the results of the sampled datasets. Tables X and XI demonstrate the statistical significance matrix based on accuracy and NMI evaluation measures.

From the statistical results, we observe that our FGB model significantly outperforms the other baseline clustering methods.

5.1.4.3 Cluster Interpretation. In order to demonstrate the capability of generating a meaningful cluster interpretation of our FGB method, a case study is conduced where we use the top 4 largest clusters of documents from the TDT2 corpora. Note that we

Table XI. The Statistical Significance Matrix Using NMI

	KM	NMF	ITCC	MSRCC	ECC	FGB
KM	-	$\ll$	$\ll$	$<$	$<$	$\ll$
NMF	$\gg$	-	$\gg$	$\gg$	$\gg$	$\ll$
ITCC	$\gg$	$\ll$	-	$\gg$	$\gg$	$\ll$
MSRCC	$>$	$\ll$	$\ll$	-	$>$	$\ll$
ECC	$>$	$\ll$	$\ll$	$<$	-	$\ll$
FGB	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	-

choose these four clusters because, (1) we want to show concrete examples (which are easier for people to read) to illustrate the ability of our method on summarization; (2) these 4 topics are well-known and familiar to readers, thus the readers can easily judge the performance of our summarization. The topics of the four document clusters are as follows: topic 1: Current Conflict with Iraq; topic 2: Monica Lewinsky Case; topic 3: 1998 Winter Olympics; and topic 4: Asian Economic Crisis. Table XII shows two-sentence summaries using our methods, and Table XIII shows the keywords selected using the NMF algorithm to describe each document cluster. Comparing the results, we observe that the summaries generated using our FGB method are more readable and they express the exact meaning of the corresponding topics. Thus, these short summaries have more powerful interpretability. We also find that many of the keywords selected by the NMF algorithm are contained in the summaries generated by our FGB method. This is because both document-term and sentence-term matrices are involved in our factorization, and the sentences containing the representative terms may have higher probabilities of being selected.

5.2 Experiments on Multidocument Summarization

5.2.1 Dataset. To evaluate the summarization results empirically, we use the DUC2002 and DUC2004 datasets for generic multidocument summarization and use DUC2005, DUC2006, and TAC2008 (set A) datasets for query-relevant summarization, both of which are open benchmark datasets from the Document Understanding Conference (DUC) for automatic summarization evaluation. Table XIV gives a brief description of the datasets for different purposes.

5.2.2 Baselines. We implement the following baseline systems to examine the quality of the summaries generated by our system.

— *LeadBase.* returns the leading sentences of all the documents for each topic, which is also used as the baseline in DUC evaluation.
— *Random.* randomly selects sentences for each topic.
— *LSA.* conducts latent semantic analysis on terms by sentences matrix as proposed in Gong and Liu [2001].
— *NMFBase.* performs NMF on terms by sentences matrix and ranks the sentences by their weighted scores.
— *KM.* calculates sentence similarity matrix using cosine similarity and performs K-means algorithm to cluster the sentences and chooses the center sentences in each cluster.
— *NMF.* similar procedures as KM and uses NMF as the clustering method.
— *DUCAve.* average scores of the DUC participants.
— *DUCBest.* the highest scores of the DUC participants.

Table XII. Two-Sentence Summaries Formed by Our FGB Method for the Four Topics

topic1	- The Security Council has refused to lift the sanctions until Iraq complies with council resolutions demanding it destroy its weapons of mass destruction. - It's an attempt to demonstrate Iraq is cooperating with U.N. monitors, and to prove it is not mass producing biological and chemical weapons.
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. - In those interviews, with a local newspaper, he described Lewinsky's visits to the Oval Office as typical transactions between an aide bearing documents and a president receiving them.
topic3	- The IOC had been expected to approve a new rule that all challenges to Olympic results must be made within three years after the games and settled by the time the next games begins. - NAGANO, Japan (AP): the Dutch created the revolutionary clap skate and would end up with the most medals and world records.
topic4	- HONG KONG (AP): southeast Asian currencies hit new lows Tuesday for a second straight day, unnerving investors and sending regional stock markets tumbling. - The fall of a HONG KONG investment company has shaken financial markets around the world.

Table XIII. Representative Terms Selected by the NMF Algorithm

topic1	government, economic, country, people, international
topic2	Clinton, president, Lewinsky, house, grand
topic3	team, games, Olympics, nagano, world
topic4	percent, market, stock, Asian, companies

— *2-Stage1.* conducts document clustering first, and then clustering sentences in each document cluster and extracts the representative sentences in each sentence cluster. K-means algorithm is used for both document clustering and sentence clustering.
— *2-Stage2.* similar procedures as 2-Stage1 while uses NMF as the clustering method.

We also compare our system with several newly proposed systems in query-relevant document summarization tasks.

— *SingleMR.* proposes a manifold-ranking based algorithm for sentence ranking [Wan et al. 2007].
— *MultiMR.* uses multimodality manifold-ranking method by utilizing within-document and cross-document sentence relationships as two separate modalities [Wan and Xiao 2009].
— *SemanSNMF.* constructs sentence similarity matrix by using semantic role analysis, and then conducts symmetric nonnegative matrix factorization on the similarity

Table XIV. Description of the Datasets Used for Multidocument Summarization

	DUC2002	DUC2004	DUC2005	DUC2006	TAC2008
Number of collections	59	50	50	50	48
Number of documents in each collection	∼10	10	25 ∼ 50	25	10
Data source	TREC	TDT	TREC	AQUAINT	AQUAINT
Summary length	100 words	665bytes	250 words	250 words	100 words
Purpose	generic	generic	query	query	query

matrix to cluster sentences, and finally selects the most important sentences in each cluster based on the within-cluster sentence selection scheme [Wang et al. 2008a].

5.2.3 Evaluation Methods. We use the ROUGE [Lin and Hovy 2003] toolkit (version 1.5.5) to measure our proposed FGB method, which is widely applied by DUC for performance evaluation. It measures the quality of a summary by counting the unit overlap between the candidate summary and a set of reference summaries. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-SU. ROUGE-N is an n-gram recall computed as follows.

$$ROUGE-N = \frac{\sum_{S\in\{ref\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{ref\}}\sum_{gram_n\in S} Count(gram_n)},$$

where n is the length of the n-gram, and ref stands for the reference summaries. Note that $Count_{match}(gram_n)$ is the maximum number of n-grams cooccurring in a candidate summary and the reference summaries, and $Count_(gram_n)$ is the number of n-grams in the reference summaries. ROUGE-L uses the longest common subsequence (LCS) statistics. Intuitively, the longer the LCS of two summaries, the more similar are the two summaries. ROUGE-W is based on weighted LCS and favors strings with consecutive matches. Each of these evaluation methods in ROUGE can generate three scores (recall, precision, and F-measure). As we have similar conclusions in terms of any of the three scores, for simplicity, in this paper we only report the average F-measure scores generated by ROUGE-1, ROUGE-2, and ROUGE-W and compare our proposed FGB method with other implemented systems. In general, the higher the ROUGE score, the more similar the two compared summaries.

5.2.4 Experimental Results.

5.2.4.1 Results of Generic Multidocument Summarization. Table XV and Table XVI show the ROUGE evaluation results on DUC2002 and DUC2004 datasets for generic multidocument summarization. We observe that our method achieves high ROUGE scores and outperforms most of the baseline systems (except the best team in DUC 2004). This is because in our FGB model, the factorization results contain both the sentence-topic matrix, from which we choose the sentences with the highest probabilities in each topic to form the summaries and the document-topic matrix, from which we can get the document clusters. Since the topics are generated from both the document side and the sentence side, the document-level and the sentence-level information will influence each other. Therefore the sentences used for document summarization are not treated independently, as do many of the existing methods. The high-level document clustering can provide a proper guide to know the context dependency of these

Table XV. The Results of Generic Multidocument Summarization Using DUC2002 Data. Remarks: * Indicates that the Improvement of Our FGB Model Over the Baseline NMF is Statistically Significant

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LeadBase	0.28684	0.05283	0.09525
Random	0.26927	0.05091	0.09358
NMFBase	0.32978	0.07252	0.10963
LSA	0.33696	0.07153	0.10827
KM	0.34127	0.07643	0.11022
NMF	0.35265	0.07867	0.11247
DUCAve	0.29411	0.05763	0.10285
FGB	**0.35626***	0.07887	**0.11734***
2-Stage1	0.34205	0.07552	0.11016
2-Stage2	0.35134	**0.07891**	0.11343
DUCBest	0.35151	0.07642	0.11448

Table XVI. The Results of Generic Multidocument Summarization Using DUC2004 Data

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LeadBase	0.32419	0.06411	0.11859
Random	0.31865	0.06377	0.11734
NMFBase	0.33179	0.06518	0.11757
LSA	0.34145	0.06538	0.12042
KM	0.34872	0.06937	0.12339
NMF	0.36747	0.07261	0.12961
DUCAve	0.33719	0.06872	0.11913
FGB	**0.38724***	0.08115*	0.13096
2-Stage1	0.35914	0.07112	0.12395
2-Stage2	0.36863	0.07201	0.12987
DUCBest	0.38224	**0.09216**	**0.13325**

sentences, which may help to improve the summarization performance. The experimental results confirm that the context information embedded in the documents do help to summarize the documents in a more meaningful way.

As seen from the results, the ROUGE scores of our methods are higher than the best team in DUC2002 and competitive with the best team from DUC2004. The good results of the best team come from the fact that they extract the topic information of the document set in an ad-hoc manner, and utilize advanced natural language processing techniques to resolve pronouns and other anaphoric expressions. However, although we also can play more on the preprocessing or language processing parts, it is not our ultimate goal. Our goal is to integrate document clustering and summarization, and the summarization experiments are set up to evaluate the quality of the summaries comparing with other existing methods using the same preprocessed data.

In order to better understand the results, we use Figure 6 and Figure 7 to visually illustrate the comparison. We subtract the LeadBase score from the scores of all the other methods in these figures so that the difference can be observed more clearly. As we have similar conclusion on different ROUGE scores, we only show the ROUGE-1 results in these figures.

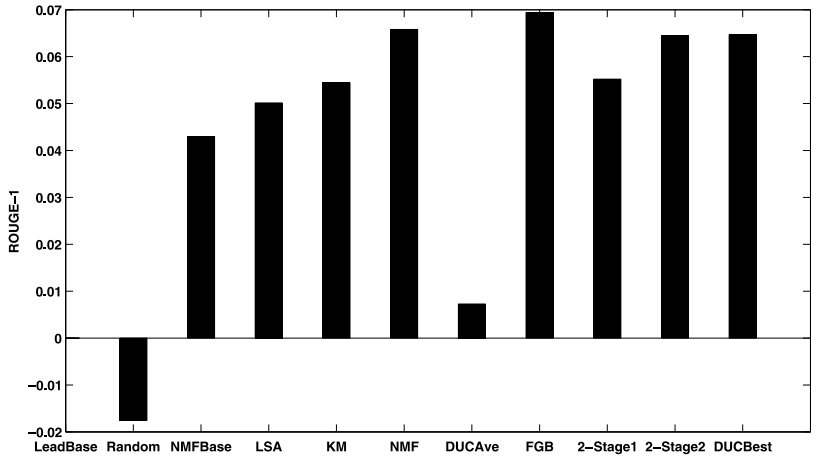

Fig. 6. Generic multidocument summarization comparison using DUC2002 data.

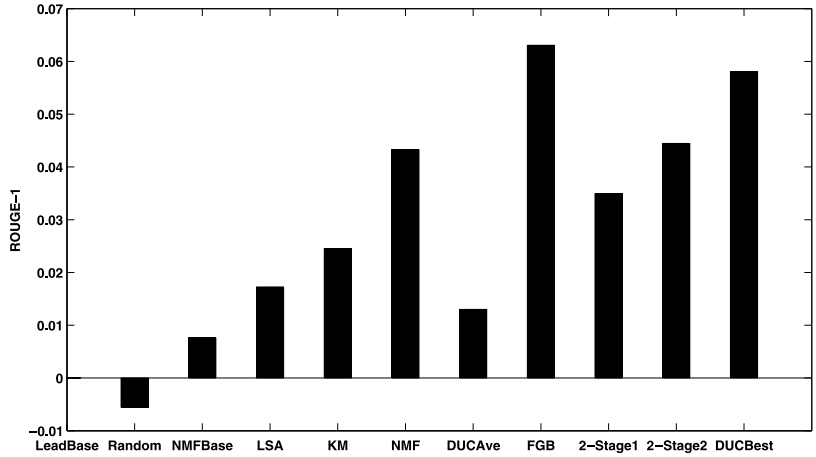

Fig. 7. Generic multidocument summarization comparison using DUC2004 data.

5.2.4.2 Results of Query-Relevant Multidocument Summarization. Sometimes people may provide specific queries to ask questions related to the content of the documents. Then the task becomes query-relevant summarization. In our framework, we select a subset of sentences that are relevant to the query. In particular, we compute the similarity between the given query and each sentence in the document collection, using the cosine similarity. Then the top 300 sentences in each document collection are selected as the candidates. Finally the selected sentence candidates instead of the entire sentence collections are used for summarization. For the baseline systems, we also use the candidate sentences as the sentence set to make them adapt to query-relevant summarization. Tables XVII–XIX show the ROUGE evaluation results on DUC2005, DUC2006, and TAC2008 (set A) datasets for query-relevant multidocument summarization. From the results, we observe the following.

— Naive methods such as Random and LeadBase provide the worst performance, as expected.

Table XVII. The Results of Query-Relevant Multidocument Summarization Using DUC2005 Data. Remark: "-" Indicates that the Method Does Not Officially Report the Results

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LeadBase	0.29243	0.04320	0.10046
Random	0.29012	0.04143	0.09802
NMFBase	0.31107	0.04932	0.10785
LSA	0.30461	0.04079	0.10883
KM	0.31762	0.04938	0.10806
NMF	0.32026	0.05105	0.11278
DUCAve	0.34347	0.06024	0.11675
FGB	**0.34851***	**0.06243***	**0.12206***
2-Stage1	0.33110	0.05105	0.11737
2-Stage2	0.33218	0.05273	0.11586
DUCBest	**0.37978**	**0.07431**	**0.12979**
SingleMR	**0.36316**	**0.06603**	**0.12694**
MultiMR	**0.36909**	**0.06836**	**0.12877**
SemanSNMF	0.35006	0.06043	0.12266

Table XVIII. The Results of Query-Relevant Multidocument Summarization Using DUC2006 Data

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LeadBase	0.32082	0.05267	0.10993
Random	0.31749	0.04892	0.10779
NMFBase	0.32374	0.05498	0.11341
LSA	0.33078	0.05022	0.11220
KM	0.33605	0.05481	0.12450
NMF	0.33850	0.05851	0.12637
DUCAve	0.37959	0.07543	0.13001
FGB	**0.38712***	**0.08295***	**0.13371***
2-Stage1	0.35115	0.06581	0.12641
2-Stage2	0.35047	0.06887	0.12973
DUCBest	**0.41017**	**0.09513**	**0.14264**
SingleMR	**0.39534**	**0.08335**	**0.13766**
MultiMR	**0.40306**	**0.08508**	**0.13997**
SemanSNMF	**0.39551**	**0.08549**	**0.13943**

— The widely used clustering-based summarization methods, e.g. KM, NMF, and LSA can improve important sentence selection.
— DUCBest in DUC2005 and DUC2006 applies natural language processing techniques to obtain a better and cleaner dataset. And the method focuses more on the topic information. The best team in TAC2008 (ID 43) [Long et al. 2009] developed a method to approximate the conditional information distance between sentences.
— Some participants in TAC2008 achieved high ROUGE scores by using syntactic parsing in the process of sentence compression, which allows systems to remove unnecessary parts of sentences, either in pre- or post-selection.

Table XIX. The Results of Query-Relevant Multidocument Summarization Using TAC2008 (Set A) Data. Remarks: "ID43," "ID13," and "ID60" are the Top Three Teams in TAC2008

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LeadBase	0.30908	0.06085	0.12062
Random	0.28754	0.05731	0.10243
NMFBase	0.32863	0.91056	0.11696
LSA	0.33125	0.90247	0.12213
KM	0.32065	0.08746	0.11367
NMF	0.33685	0.09512	0.13011
FGB	**0.35981**	**0.09973***	**0.14938***
2-Stage1	0.34527	0.09642	0.13825
2-Stage2	0.34689	0.097731	0.14012
ID13	**0.37461**	**0.10915**	**0.15174**
ID60	**0.38223**	**0.10116**	**0.15372**
ID43	**0.37696**	**0.10943**	**0.15724**

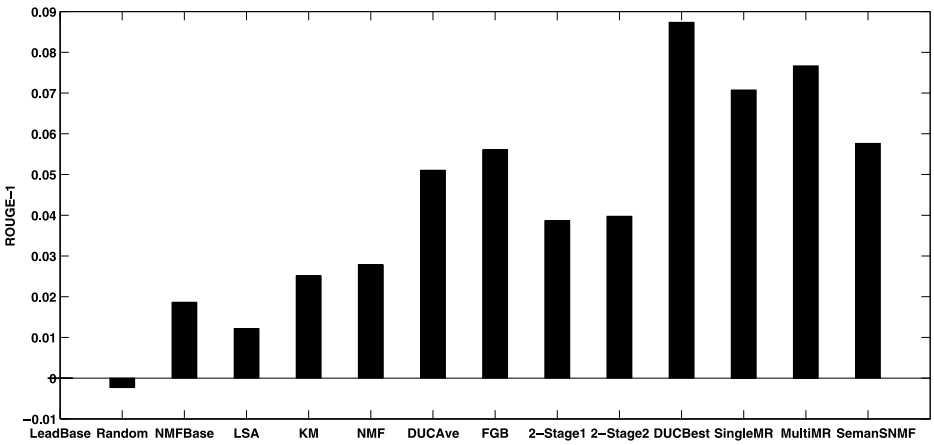

Fig. 8. Query-relevant multidocument summarization comparison using DUC2005 data.

— The new methods proposed in recent years greatly improve the summarization results by using various advanced techniques such as semantic analysis and manifold ranking.
— Our method achieves high ROUGE scores and outperforms most of the baseline systems, and is comparable with newly developed summarizers and the best DUC participant.

Note that the goal of our work is not building a query-based summarization system. We aim to provide an integrated system to cluster documents in the while summarizing each document cluster for users to better understand documents. Figures 8–10 visually illustrate the comparison.

6. CONCLUSION AND FUTURE WORK

In this article, we propose a new method to integrate document clustering and multidocument summarization to improve document understanding. This method

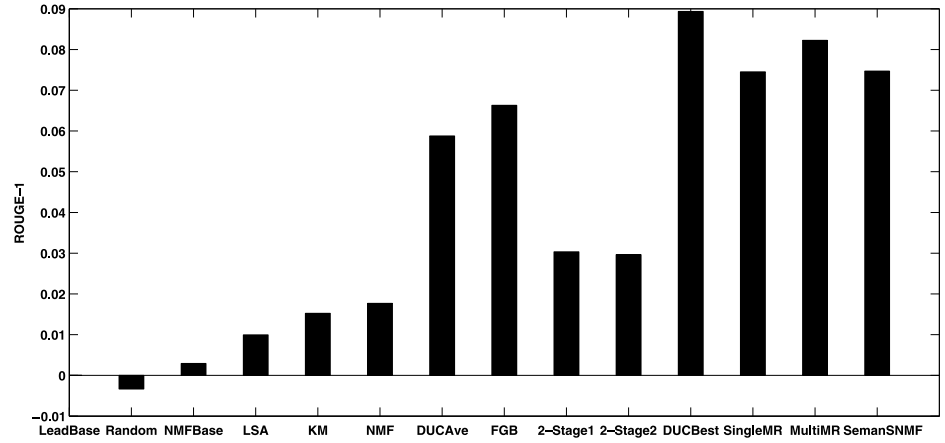

Fig. 9. Query-relevant multidocument summarization comparison using DUC2006 data.

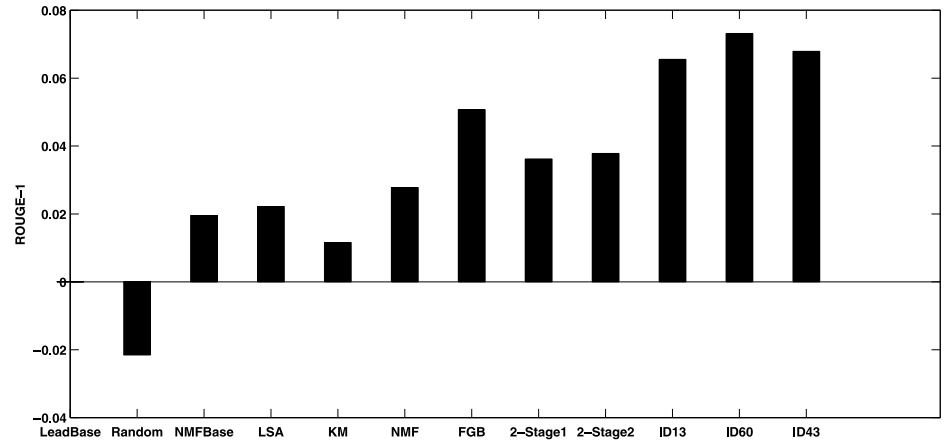

Fig. 10. Query-relevant multidocument summarization comparison using TAC2008 (Set A) data.

considers a model that generates documents as a mixture of clusters, which in turn are mixtures of base language models. By estimating the parameters, we derive the probability of documents and sentences, given clusters, which gives the scores to select the summary sentences for each document cluster. Thus, we can simultaneously cluster the documents and interpret the document clusters using the selected sentences as the summaries. The experiments on both the clustering and multidocument summarization sides using TDT2, Reuters, and DUC datasets show the effectiveness of our method.

For future work, we will improve the following aspects.

(1) Instead of using the entire sentence collection, we will explore how to find more meaningful sentence subsets as the given bases. For different tasks, the sentence subset may be selected based on different criteria. For example, in the query-relevant summarization experiments, we select the sentences similar to the queries as the candidates, and in the opinion clustering and summarization, we may use sentences expressing positive/negative opinions as our candidates, and so on.

(2) We will conduct user studies on more datasets to validate the robustness of our methods. Since we do not have the ground truth for most of the large-scale data, such as blog data, performing the user study is the best way to compare and evaluate the clustering and summarization results.

REFERENCES

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2002. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani Eds. MIT Press, Cambridge, MA, 601–608.

CHO, H., DHILLON, I., GUAN, Y., AND SRA, S. 2004. Minimum sum squared residue co-clustering of gene expression data. In *Proceedings of SIAM International Conference on Data Mining*.

CONROY, J. AND O'LEARY, D. 2001. Text summarization via hidden Markov models. In *Proceedings of SIGIR*. 406–407.

DEVORE, J. AND PECK, R. 1977. *Statistics: The Exploration and Analysis of Data*. Duxbury Press.

DHILLON, I. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of ACM (SIGKDD) International Conference on Knowledge Discovery and Data Mining*. 269–274.

DHILLON, I., MALLELA, S., AND MODHA, S. 2001. Information-theoretic co-clustering. In *Proceedings of ACM (SIGKDD) International Conference on Knowledge Discovery and Data Mining*. 89–98.

DING, C., HE, X., ZHA, H., GU, M., AND SIMON, H. 2001. A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. 107–114.

DING, C., LI, T., PENG, W., AND PARK, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of ACM (SIGKDD) International Conference on Knowledge Discovery and Data Mining*. 126–135.

DUDA, R., HART, P., AND STORK, D. 2001. *Pattern Classification*. John Wiley and Sons, Inc.

DUNLAVY, D., O'LEARY, D., CONROY, J., AND SCHLESINGER, J. 2007. QCS: A system for querying, clustering and summarizing documents. *Inform. Process. Manag. Int. J.*

ELKAN, C. 2006. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of International Conference on Machine Learning (ICML)*. 289–296.

ERKAN, G. AND RADEV, D. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of International Conference on Empirical Method on Natural Language Processing (EMNLP)*.

GOLDSTEIN, J., KANTROWITZ, M., MITTAL, V., AND CARBONELL, J. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the International ACM SIGIR Conference on Research and Development on Information Retrieval*. 121–128.

GONG, Y. AND LIU, X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the International ACM SIGIR Conference on Research and Development on Information Retrieval*. 75–95.

HE, J., LAN, M., TAN, C., SUNG, S., AND LOW, H. 2004. Initialization of cluster reffinement algorithms: A review and comparative study. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*.

HOFFMAN, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development on Information Retrieval*.

JING, H. AND MCKEOWN, K. 2000. Cut and paste based text summarization. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

JING, L., NG, M. K., AND HUANG, J. Z. 2007. An entropy weighting k-means algorithm for subspace clustering of high dimensional sparse data. *IEEE Trans. Knowl. Data Eng.* 1026–1041.

KNIGHT, K. AND MARCU, D. 2002. Summarization beyond sentence extraction: A probablistic approach to sentence compression. *Artif. Intell.* 91–107.

LEE, D. D. AND SEUNG, H. S. 2001. Algorithms for non-negative matrix factorization. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*.

LI, T. 2005. A general model for clustering binary data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 188–197.

LI, T. AND DING, C. 2006. The relationships among various nonnegative matrix factorization methods for clustering. In *Proceedings of the IEEE International Conference on Data Mining*. 362–371.

LI, T., MA, S., AND OGIHARA, M. 2004. Document clustering via adaptive subspace iteration. In *Proceedings of the International ACM SIGIR Conference on Research and Development on Information Retrieval*. 218–225.

LIN, C.-Y. AND HOVY, E. 2001. From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of Association for Computational Linguistics (ACL)*. 457–464.

LIN, C.-Y. AND HOVY, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NLT-NAACL)*. 71–78.

LIU, X., GONG, Y., XU, W., AND ZHU, S. 2003. Document clustering with cluster refinement and model selection capabilities. In *Proceedings of the International ACM SIGIR Conference on Research and Development on Information Retrieval*. 191–198.

LONG, B., WU, X., ZHANG, Z. M., AND YU, P. S. 2006. Unsupervised learning on k-partite graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 317–326.

LONG, C., HUANG, M., ZHU, X., AND LI, M. 2009. Multi-document summarization by information distance. In *Proceedings of International Conference on Data Mining (ICDM)*.

MANA-LOPEZ, M. J., BUENAGA, M. D., AND GOMEZ-HIDALGO, J. M. 2004. Multidocument summarization: An added value to clustering in interactive retrieval. *ACM Trans. Inform. Syst*.

MANI, I. 2001. *Automatic Summarization*. John Benjamins Publishing Company.

MCKEOWN, K. R., BARZILAY, R., EVANS, D., HATZIVASSILOGLOU, V., KLAVANS, J. L., NENKOVA, A., SABLE, C., SCHIFFMAN, B., AND SIGELMAN, S. 2002. Tracking and summarizing news on a daily basis with Columbia's newsblaster. In *Proceedings of the 2nd International Conference on Human Language Technology Research*.

MIHALCEA, R. AND TARAU, P. 2005. A language independent algorithm for single and multiple document summarization. In *Proceedings of International Conference on Natural Language Processing (IJCNLP)*.

NASTASE, V. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of Conference on Empirical Methods on Natural Language Processing (EMNLP)*. 763–772.

PARK, S., LEE, J.-H., KIM, D.-H., AND AHN, C.-M. 2007. Multi-document summarization based on cluster using non-negtive matrix factorization. In *Proceedings of Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*.

RADEV, D., JING, H., STYS, M., AND TAM, D. 2004. Centroid-based summarization of multiple documents. *Inform. Process. Manag.*, 919–938.

RICARDO, B. AND BERTHIER, R. 1999. *Modern Information Retrieval*. ACM Press.

SHEN, D., SUN, J.-T., LI, H., YANG, Q., AND CHEN, Z. 2007. Document summarization using conditional random fields. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. 2862–2867.

SHI, J. AND MALIK, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.* 888–905.

STREHL, A. AND GHOSH, J. 2003. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 583–617.

TANG, J., YAO, L., AND CHEN, D. 2009. Multi-topic based query-oriented Summarization. In *Proceedings of SIAM International Conference on Data Mining (SDM)*.

TURPIN, A., TSEGAY, Y., HAWKING, D., AND WILLIAMS, H. 2007. Fast generation of result snippets in Web search. In *Prodeedings of ACM SIGIR Conference on Research and Development on Information Retrieval*. 127–134.

WAN, X. AND XIAO, J. 2009. Graph-based multi-modality learning for topic-focused multi-document summarization. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. 1586–1591.

WAN, X. AND YANG, J. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st ACM Annual International SIGIR Conference on Research and Development on Information Retrieval*.

WAN, X., YANG, J., AND XIAO, J. 2007. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. 2903–2908.

WANG, D., LI, T., ZHU, S., AND DING, C. 2008a. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of ACM SIGIR Conference on Research and Development on Information Retrieval*.

WANG, D., ZHU, S., LI, T., CHI, Y., AND GONG, Y. 2008b. Integrating clustering and multi-document summarization to improve document understanding. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*. 1435–1436.

WANG, D., ZHU, S., LI, T., AND GONG, Y. 2009. Multi-document summarization using sentence-based topic models. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL'09)*.

WANG, F., ZHANG, C., AND LI, T. 2007. Regularized clustering for documents. In *Proceedings of ACM SIGIR Conference on Research and Development on Information Retrieval*. 95–102.

WEI, F., LI, W., LU, Q., AND HE, Y. 2008. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*. ACM, 283–290.

XU, W. AND GONG, Y. 2004. Document clustering by concept factorization. In *Proceedings of ACM SIGIR Conference on Research and Development on Information Retrieval*. 202–209.

XU, W., LIU, X., AND GONG, Y. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of ACM SIGIR Conference on Research and Development on Information Retrieval*. 373–386.

YIH, W.-T., GOODMAN, J., VANDERWENDE, L., AND SUZUKI, H. 2007. Multi-document summarization by maximizing informative content-words. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. 1776–1782.

ZAMIR, O. AND ETZIONI, O. 1998. Web document clustering: A feasibility demonstratio. In *Proceedings of ACM SIGIR Conference on Research and Development on Information Retrieval*. 46–54.

ZHA, H., HE, X., DING, C., GU, M., AND SIMON, H. 2001. Bipartite graph partitioning and data clustering. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*. 25–32.

ZHONG, S. AND GHOSH, J. 2003. A unified framework for model-based clustering. *J. Mach. Learn. Res.*, 1001–1037.

ZHONG, S. AND GHOSH, J. 2005. Generative model-based document clustering: A comparative study. *Knowl. Inf. Syst.*, 374–384.

ZIEN, J., SCHLAG, M., AND CHAN, P. K. 1999. Multilevel spectral hypergraph partitioning with artibary vertex sizes. *IEEE Trans. Comput.-Aid. Design*. 1389–1399.

Received June 2009; revised February 2010; accepted July 2010