JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 32, 1613-1634 (2016)

DOI:10.1688/JISE.2016.32.6.12

Multi-document Summarization using Probabilistic Topic-based Network Models*

CHENG-ZEN YANG, JHIH-SHANG FAN AND YU-FAN LIU
Department of Computer Science and Engineering
Yuan Ze University
Chungli, 32003 Taiwan
E-mail: czyang@syslab.cse.yzu.edu.tw; {s1003305, s1001447}@mail.yzu.edu.tw

Multi-document summarization has obtained much attention in the research domain of text summarization. In the past, probabilistic topic models and network models have been leveraged to generate summaries. However, previous studies do not investigate different combinations of various topic models and network models. This paper describes an integrated approach considering both probabilistic topic models and network models. Two probabilistic topic models and four network models are investigated. We have conducted experiments to evaluate the effectiveness of the proposed approach with the DUC 2004-2007 datasets and make a systematic comparison between two representative topic models, PLSA and LDA. The results show that the PLSA-based network approach outperforms the TF-IDF baseline on all datasets. Moreover, PLSA has better ROUGE performance than LDA for multi-document summarization.

Keywords: multi-document summarization, probabilistic topic models, network models, extraction-based summarization, performance evaluation

1. INTRODUCTION

Automatic multi-document summarization is a challenging problem that has otained significant attention in the research domain of text summarization [1-4]. Given a collection of related documents, the goal of multi-document summarization is to generate a concise summary containing important information as much as possible.

The approaches of multi-document summarization can be mainly categorized into two classes: the abstractive approach and the extractive approach [5, 6]. The abstractive approach produces summaries that are paraphrased from the source documents. Most abstractive methods are knowledge-rich methods requiring abundant support from natural language processing and domain-specific ontologies [1]. On the contrary, the extractive approach generates a summary by selecting a subset of informative sentences from the source documents. Heuristic rules or learning models are leveraged to decide the importance of the sentences.

Although the output of the abstractive approach is much closer to the manual summary by human, the extractive approach has shown its prominence in multi-document summarization [5]. Recently, many extractive methods have been proposed [2]. For example, a template-based method has been developed in SUMMONS [7, 8]. A cluster centroid-based method MEAD is proposed to compute the thematic importance of the sentences [9]. Graph-based methods have been investigated in various research studies, such as the cohesion-based approach [10], the affinity graph approach [11], LexRank [5],

Received August 27, 2015; revised October 19, 2015; accepted December 9, 2015.
Communicated by Hsin-Hsi Chen.
* This work was supported in part by Ministry of Science and Technology, Taiwan, under Grant No. MOST 104-221-E-155-004.

and iSpreadRank [12]. Some studies deal with this problem as an optimization problem of selecting informative sentences using metaheuristic algorithms, such as the differential evolution (DE) approach [13].

Recently, network models have demonstrated their effectiveness in multi-document summarization [5, 12, 14-16]. Moreover, Latent topic models have been used to improve the summarization performance. A well-known Latent Dirichlet Allocation (LDA) model [17] has been discussed in many studies, *e.g.*, [14, 15, 18-20]. Another topic model Probabilistic Latent Sematic Analysis (PLSA) [21] has also been discussed, *e.g.*, [22]. As shown in [23, 24], both can achieve comparable performance in different tasks. However, these two models have their own shortcomings: the performance of LDA highly depends on its hyper-parameter settings and PLSA has the overfitting problem [23]. These studies motivate us to investigate the effectiveness of both PLSA and LDA with the network models in multi-document summarization.

In this paper, we propose an extractive approach using probabilistic topic-based network models for multi-document summarization. Two representative topic models PLSA and LDA are investigated in deriving the latent topics of sentences. Then, four network models are explored in calculating the rank of the sentences according to their latent topic features. Finally, we adopt the CSIS (Cross-Sentence Information Subsumption) [9] approach to reduce the semantic redundancy for summary generation.

We have conducted experiments on the datasets of DUC 2004-2007 (Document Understanding Conferences). The results show that the proposed extractive approach can have high performance in most cases. The three main contributions of this work are:

(1) A multi-document summarization approach based on the probabilistic topic-based network models is proposed. Two probabilistic topic models (PLSA and LDA) and four network models (Degree centrality, Normalized Similarity-based Degree centrality, PageRank, and iSpreadRank) are investigated.
(2) To the best of our knowledge, there has been no systematic comparison between PLSA and LDA for the multi-document summarization task. This work investigates their performance with the DUC 2004-2007 datasets. The results show that PLSA outperforms LDA by effectively capturing salient topics.
(3) Comprehensive experimental studies are conducted with four DUC datasets. Compared with other state-of-the-art approaches on all datasets, the PLSA-based network approach can stably have high performance.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. Section 3 describes the details of the probabilistic topic models and the network models in the proposed approach. Section 4 presents the empirical results of the proposed approach and the comparisons with previous work. Finally, Section 5 concludes this paper.

2. RELATED WORK

Generic multi-document summarization has acquired much attention in the summarization research field. A variety of extractive approaches exist for selecting the most salient sentences from a collection of topic-related documents [2]. In this section, we only review the previous studies that use the following two models: probabilistic topic models and network models. More information about other related multi-document summarization approaches can be found in the related survey papers [2, 3].

2.1 Probabilistic Topic-based Approaches

For multi-document summarization, topicality of sentences has been explored in many past studies, such as [25, 26]. Recently, many approaches are devised by leveraging modern probabilistic topic models derived from the Latent Semantic Index (LSI) model [27]. LSI is effective for information retrieval tasks because it uses singular value decomposition (SVD) to extract the latent semantics of documents by mapping a high-dimensional word-document matrix to a low-dimensional sematic space. However, LSI has significant computational overhead for SVD [28]. Moreover, LSI has statistical shortages because of the implicit Gaussian noise assumption for term frequencies [21].

The Probabilistic Latent Sematic Analysis (PLSA) model [21] is proposed subsequently. In PLSA, a document is composed of latent topics, and each word w_i has a topic-specific word distribution $P(w_i|z_k)$ associated with latent topic z_k. Based on the maximization of the likelihood measures, it has better performance in extracting latent topic semantics. Recently, Hennig proposed a PLSA-based approach for query-focused multi-document summarization by extracting thematic features from queries, document titles, and narratives [29]. The experimental results show that PLSA can achieve outstanding ROUGE performance.

The Latent Dirichlet Allocation (LDA) model is a generative probabilistic model [17]. Its Bayesian hierarchy consists of three levels: the word level, the document level, and the corpus level. Compared with PLSA, LDA introduces the Dirichlet priors to model the document-specific topic distributions and topic-specific word distributions. Therefore, it can be used to model the latent topic space for unseen documents. In [18], Arora and Ravindran propose an LDA-based approach for multi-document summarization. However, their approach assumes each sentence belongs to only one topic.

In [19], a hybrid scheme is proposed using a hierarchical LDA model to extract sentence topics and then a supervised learning model to generate rank scores for sentences. In this scheme, however, the inherited weakness of the supervised learning may limit the summarization performance while processing new documents with unseen topics.

Xu, Liu, and Araki propose a hybrid topic model for multi-document summarization using the Hidden Topic Markov Model (HTMM) [30] to extract topics and decide the binary topic transition relationships with a surface texture model [31]. Then the topic transition model is leveraged to re-rank the sentences by considering their probability transitions. However, HTMM only considers local dependencies among topics. A new sentence can either continue the old topic or switch to a new topic. The global dependency is not considered. In addition, sentences are not allowed to have any topic transition.

2.2 Network-based Approaches

Due to the emerging development of network analysis techniques for Web, many multi-document summarization approaches leverage network models to rank the sentences. For example, Erkan and Radev propose a graph-based approach called LexRank incorporating the PageRank model [32] to calculate the sentence salience to the latent topics [5]. Weighted cosine similarity graphs are constructed according to the similarity measures of sentences. LexRank then computes the ranking score of a sentence by considering the similarity influences of its adjacent sentences. Mihalcea and Tarau further study the effectiveness of two Web ranking models, HITS [33] and PageRank, in a two-layer summarization framework for multi-document summarization [34]. They find that the layered framework of network models has very competitive summarization performance to the state-of-the-art summarization systems.

iSpreadRank adopts the Leaky Capacitor Model [35] to iteratively consider the spreading influences of the neighbor nodes in the graph [12]. As shown in [16], iSpreadRank centrality can get performance improvements over HITS and PageRank.

Many link analysis algorithms, such as PageRank, can be illustrated as a Markov Random Walk model. In [36], Wu and Yang propose two graph-based models by leveraging the cluster information in the Conditional Markov Random Walk (ClusterCMRW) [37] model and the HITS algorithm (ClusterHITS). In ClusterCMRW and ClusterHITS, a two-layer link graph is constructed to employ the information of theme clusters produced by a clustering scheme. Three clustering algorithms are discussed: *K*-means, Agglomerative clustering and Divisive clustering. Based on the DUC 2001-2002 datasets, both models can outperform the baseline MRW model. In [38], Fukumoto *et al*. improve the performance of ClusterCMRW by using the Spectral clustering algorithm on an NTCIR-3 dataset. In [39], Wu and Zhang propose CTSUM by leveraging the certainty information in a graph-based model. CTSUM outperforms ClusterHITS for the DUC 2007 dataset.

2.3 Topic-based Network Approaches

In the past, several studies consider both topic models and network models for multi-document summarization. For ease of understanding the following literature review, S, Z, and W represent the sentences, the topics, and the words in the sentences, respectively.

In [14], Gao *et al*. propose a topic-sentence bipartite graph approach in which the edges from sentences to topics represent the per-topic distributions $P(Z|S)$ and the edges from topics to sentences are modeled with the average of word distributions $P(W|Z)$. They use LDA to derive these distributions and HITS to calculate the salience scores of sentences. With the mutual reinforcement process of HITS, the importance scores of the sentences are adjusted according to the iteratively propagated influence scores of the topics. However, the per-sentence distribution $P(S|Z)$ of each topic is calculated by approximating it with the average word distributions in the bipartite graph assuming that words are independent. Therefore, the influences of contextual correlations among words are neglected in this model. Moreover, a sentence with more common words will obtain a relatively large $P(S|Z)$.

Pei *et al*. propose two topic-oriented network models, ToHITS and ToPageRank, to derive the salience rank of sentences [15]. ToHITS is similar to the topic-sentence bipartite graph approach of [14], but it only uses the average of word distributions $P(W|Z)$ as the per-topic distributions $P(S|Z)$ to model all edge weights. The influences of per-topic distributions $P(Z|S)$ of sentences are not considered.

ToPageRank first leverages the Topical PageRank model [20] to adjust the PageRank score of each sentence on each topic by considering per-sentence distribution $P(S|Z)$ in the random jump calculation, and then calculates the salience score of the sentence by summing up all its PageRank scores on difference topics with document-based topic weighting. Since the average of word distributions $P(W|Z)$ is also used to approximate $P(S|Z)$ in this model, the approximation has the same issues as the work of [14]. In addition, as the number of topics increases, ToPageRank needs more computation resources to perform PageRank-like computations for each latent topic.

3. SUMMARIZATION APPROACH

This section describes the details of the proposed approach. The processing flow is first briefly overviewed. Then the topic-based representation is presented. Finally, different network models are described.

3.1 Multi-document Summarization Process

Fig. 1 illustrates the processing flow. All documents are first processed with generic text pre-processing techniques, such as tokenization, stop-word removal, and stemming, to extract feature vectors. The feature vector for sentence s_j is represented in the bag-of-words model as $s_j = \langle tf_{1,j}, tf_{2,j}, \ldots, tf_{n,j} \rangle$ where $tf_{i,j}$ is the term frequency of the ith term w_i in the sentence s_j. All these sentence feature vectors are included in a term-sentence matrix TS for the following computation to extract topic-sentence relationships.

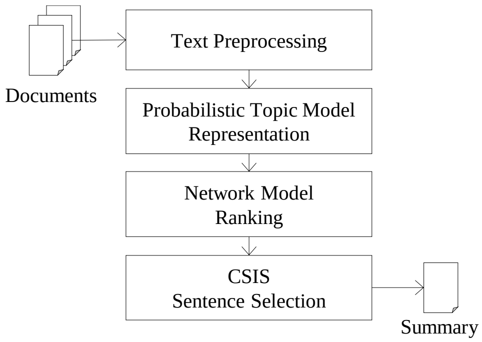


Fig. 1. The processing flow for summary generation.

The probabilistic topic model is then used to calculate the probability distribution $P(z_k|s_j)$ for a latent topic z_k given sentence s_j. For sentence s_j, its topic-based feature vector is thus $\langle P(z_1|s_j), P(z_2|s_j), \ldots, P(z_K|s_j) \rangle$. In this work, we investigate two basic representative topic models, PLSA and LDA. These topic-based vectors are then used to calculate the topic-based connection relationships among the sentences. Fig. 2 illustrates the topic-based representation of the sentences in the documents. Each sentence is extracted from the document and represented as a vector of topic-based features. Finally, network models are employed to calculate the sentence scores according to these topic-based connection relationships. In this work, four network models are investigated: Degree centrality, Normalized Similarity-based Degree centrality, PageRank, and iSpreadRank.

To generate the summary, CSIS (Cross-Sentence Information Subsumption) [9] is used to reduce the semantic redundancy. All semantically redundant candidate sentences are omitted in the summary generation process of CSIS.

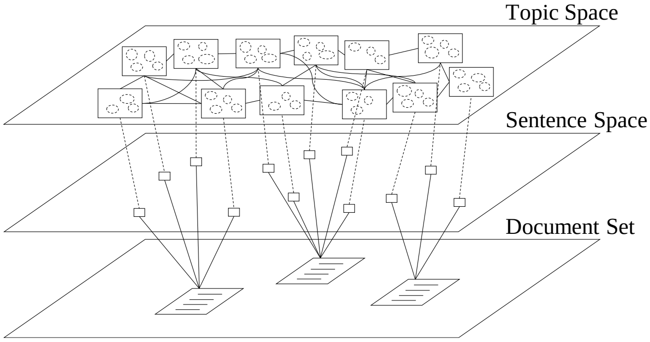


Fig. 2. The topic-based representation of the sentences in the documents.

3.2 Document Preprocessing

In the proposed approach, each sentence is first converted into the corresponding feature vector $s_j = \langle tf_{1,j}, tf_{2,j}, \ldots, tf_{n,j} \rangle$ following the generic preprocessing steps: tokenization, stop-word removal, and stemming. In the tokenization step, only words consisting of al- phanumeric characters are kept. Then, the words in the stop-word corpus are removed. The stop-word corpus is based on a public Onix stop-word list[1]. The remaining words are then stemmed using Porter stemmer.

To avoid including sentences that are very short and unlikely included in the summary, a threshold L_T is used to discard the sentences having less than L_T words as [29]. The term frequency *tf* scores of the words are then computed for remaining sentences.

For the document corpus D of the topic related documents, its corresponding term-sentence matrix TS contains all remaining sentences of D in which $TS_{i,j}$ is the term frequency of the ith term w_i in sentence s_j. Given a topic number K, the document-specific topic distributions are derived from TS using the probability topic model.

3.3 Latent Topic Extraction

In this work, we study two probabilistic topic models: PLSA [21] and LDA [17]. They are two representative topic models achieving comparable performance in different tasks [23, 24]. However, each has its own shortcomings. Based on these previous studies, we investigate their effectiveness in the proposed network models.

3.3.1 PLSA model

In the original document-based PLSA model, a latent topic space $Z=\{z_1,..., z_K\}$ is introduced to calculate the co-occurrence distribution $P(w_i, d_j)$ of a word $w_i \in W=\{w_1,..., w_M\}$ and a document $d_j \in D=\{d_1,..., d_N\}$. With the latent topic space Z, the joint probability $P(w_i, d_j)$ can be calculated as follows:

$$P(w_i, d_j) = P(d_j)P(w_i \mid d_j) = P(d_j) \sum_{\forall z_k \in Z} P(w_i \mid z_k)P(z_k \mid d_j) , \tag{1}$$

where $z_k \in Z$ is an unobserved topic, $P(w_i|z_k)$ is the topic-specific word distribution, and

[1] http://www.lextek.com/manuals/onix/stopwords1.html

$P(z_k|d_j)$ is the document-specific topic distribution. To determine $P(z_k|d_j)$ and $P(w_i|z_k)$ in Eq. (1), the log-likelihood function

$$L = \sum_{d\in D}\sum_{w\in W} n(d,w)\log P(d,w) \tag{2}$$

is maximized using the Expectation Maximization (EM) algorithm, where $n(d,w)$ is the number of times w occurred in d. In the E-step, the posterior probability for the latent topic z_k can be derived from the current estimates of the parameters:

$$P(z_k \mid w_i, d_j) = \frac{P(w_i \mid z_k)P(z_k \mid d_j)}{\sum_{l=1}^{K} P(w_i \mid z_l)P(z_l \mid d_j)}\ . \tag{3}$$

In the M-step, $P(w_i|z_k)$ and $P(z_k|d_j)$ are updated using the following equations:

$$P(w_i \mid z_k) = \frac{\sum_{j=1}^{N} n(d_j, w_i)P(z_k \mid w_i, d_j)}{\sum_{m=1}^{M}\sum_{j=1}^{N} n(d_j, w_m)P(z_k \mid w_m, d_j)}\ , \tag{4}$$

$$P(z_k \mid d_j) = \frac{\sum_{m=1}^{M} n(d_j, w_m)P(z_k \mid w_m, d_j)}{\sum_{l=1}^{K}\sum_{m=1}^{M} n(d_j, w_m)P(z_l \mid w_m, d_j)}\ . \tag{5}$$

The alternating iteration of the E-step and the M-step is a convergent procedure to approach a local maximum of the log-likelihood Eq. (2). This work uses PLSA to calculate the topic-based vector of sentence s_j as $P(z|s_j)=\langle P(z_1|s_j), P(z_2|s_j),\ldots, P(z_K|s_j)\rangle$ by Eq. (5).

3.3.2 LDA model

Compared with PLSA, the Latent Dirichlet Allocation (LDA) model instead uses a conjugate Dirichlet prior to provide prior observations for topic z_k sampled in a document [17]. Both $P(Z|S)$ and $P(W|Z)$ are modeled with the Dirichlet priors θ and ϕ, and hyper-parameters α and β of Dirichlet priors are introduced for $P(Z|S)$ and $P(W|Z)$, where θ_i is the topic distribution of document d_i, and ϕ_k is the word distribution of topic z_k. In the original LDA model, the joint probability of a topic mixture θ, a set of M topics Z, and a set of M words W is expressed as:

$$P(\theta, Z, W \mid \alpha, \beta) = P(\theta|\alpha)\prod_{m=1}^{M} P(z_m \mid \theta)P(w_m|z_m, \beta). \tag{6}$$

Since the computation of the original LDA model is complicated, several approximate inference techniques can be used to speed up the computation. One commonly used approximate is Gibbs sampling [40]. After the estimation, the topic-document distribution $P(z_k|d_j)$ can be estimated with the Gibbs sample as:

$$P(z_k \mid d_j) = \frac{n(d_j|\ z_k) + \alpha}{\sum_k n(d_j|\ z_k) + K\alpha}, \tag{7}$$

where $n(d_j|z_k)$ is the number of words in document d_j that have been assigned to topic z_k. In this work, we investigate the effectiveness of LDA by using Eq. (7) to alternatively calculate the LDA-based topic feature vector of sentence s_j.

3.4 Topic-based Network Models

With the probabilistic topic models, all sentences have their own topic-based feature vectors. A topic-aspect network can be constructed to express the topic relationship of these sentences. We follow the similar hypothesis as addressed in LexRank and iSpread-Rank but in respect to the latent topic space: the sentences are said to be more salient in the latent topic space when their topic feature vectors are similar to many topic feature vectors of the other sentences. Therefore, network models are used to calculate the topic centrality score of each sentence.

In the following sections, we first describe the network construction in the latent topic space. We thereafter discuss four network models for centrality computation.

3.4.1 Topic-similarity graph construction

Based on the extracted topic feature vectors, a complete topic-based graph $G=(V, E)$ can be constructed in which nodes in $V=S=\{s_1,\ldots, s_N\}$ are the topic feature vectors of the sentences and edges in E represent the relationships between a pair of sentences. However, from the aspect of topic-similarity, some edges can be ignored because their similarity values are less significant.

In this work, two kinds of similarity are considered separately to decide the existence of the edges. The first is the cosine similarity. For two sentences s_{j1} and s_{j2}, the topic-aspect cosine similarity is defined as:

$$sim_{\text{cosine}}(s_{j1}, s_{j2}) = \frac{P(z \mid s_{j1}) \cdot P(z \mid s_{j2})}{|P(z \mid s_{j1})| \times |P(z \mid s_{j2})|}, \tag{8}$$

where $P(z|s_{j1})$ and $P(z|s_{j2})$ are the topic-based feature vectors of sentence s_{j1} and s_{j2}.

The second considered similarity is the Jensen-Shannon (JS) divergence (D_{JS}). Since JS-divergence is a symmetrized and smoothed version of the Kullback-Leibler (KL) divergence (D_{KL}), the KL divergence is not discussed in this work. The JS-divergence similarity of two sentences s_{j1} and s_{j2} is defined as:

$$\begin{aligned} sim_{\text{JS}}(s_{j1}, s_{j2}) &= 1 - D_{JS}(s_{j1} \,\|\, s_{j2}) \\ &= 1 - [\tfrac{1}{2} D_{KL}(s_{j1} \,\|\, s_{jm}) + \tfrac{1}{2} D_{KL}(s_{j2} \,\|\, s_{jm})], \end{aligned} \tag{9}$$

where $s_{jm}=1/2(s_{j1}+s_{j2})$ and the D_{KL} is defined as:

$$D_{KL}(s_{j1} \,\|\, s_{j2}) = \sum_k P(z_k \mid s_{j1}) \log \frac{P(z_k \mid s_{j1})}{P(z_k \mid s_{j2})}. \tag{10}$$

These similarities are then ranked in decreasing order, and a threshold γ is used to trim off the bottom γ% of edges with small similarities. If there are nodes having no connected edges, these *isolated nodes* are removed, because the sentences represented by these nodes are potentially irrelevant to other sentences. Therefore, we can get two topic-similarity graphs G_{cos} and G_{JS} for the following centrality computation.

3.4.2 Centrality computation

This work is similar as the previous study in [16] to discuss the network models for sentence centrality computation: degree centrality, normalized Similarity-based degree centrality, PageRank centrality, and iSpreadRank centrality. However, there are two major differences between this previous study and our work. First, these network models are leveraged in this work to compute the sentence salience for the topic-similarity graphs, but the previous study discusses these network models for the sentence-similarity graphs. Second, this work does not discuss HITS centrality because of its poor performance in the previous study.

3.4.2.1 Degree centrality

In a topic-similarity graph like G_{cos} or G_{JS}, the *degree centrality* score ($DC(s_j)$) of sentence s_j is defined as the degree of the corresponding node in the topic-similarity graph (*i.e.*, the topic vector $P(z|s_j)$). A sentence has a high degree centrality score when the latent topics of this sentence are similar to the latent topics of many other sentences.

This work does not discuss *weighted degree centrality* as studied in [16] because the weighted degree centrality is only an extension of degree centrality by considering the similarity weights of connected edges, not just the edge number. For the purpose of demonstrating the performance difference between the baseline network model and other advanced network models, using plain degree centrality can fulfill this purpose.

3.4.2.2 Normalized similarity-based degree centrality

In the aforementioned definition of degree centrality, one obvious drawback is that this centrality does not consider the influences of the topic similarity. If the number of the connected nodes of a sentence is the same as another sentence, these two sentences have the same degree centrality score.

To cope with this drawback, the topic-similarity scores can be leveraged as the similarity weights of the edges. In addition, the similarity weight of an edge is normalized by considering the total similarity weight of the connected neighbor node. The normalized similarity-based degree centrality score ($DC_{NS}(s_j)$) of sentence s_j is thus defined as:

$$DC_{NS}(s_j) = \sum_{e_{i,j} \in E} \frac{w(e_{i,j})}{\sum_{e_{i,k} \in E} w(e_{i,k})}, \tag{11}$$

where $w(e_{i,j})$ is the topic-similarity weight of edge $e_{i,j}$. In this work, two topic-similarity graphs G_{cos} and G_{JS} are investigated.

3.4.2.3 Pagerank centrality

PageRank is a random walk model originally used to rank the Web search results [32].

In this work, we leverage PageRank to consider the semantic influences among sentences in the topic-similarity graphs. In this work, PageRank is applied to the undirected topic-similarity graphs because the cosine similarity and JS-divergence are all symmetric.

The PageRank centrality score ($PR(s_j)$) of sentence s_j is thus defined as:

$$PR(s_j) = \frac{d}{N} + (1-d) \sum_{\forall k: e_{j,k} \in E} \frac{PR(s_k)}{\deg(s_k)}, \tag{12}$$

where d is the damping factor which is typically 0.15, and $\deg(s_k)$ is the degree of s_k.

3.4.2.4 iSpreadrank centrality

iSpreadRank is a graph-based ranking mechanism to determine the sentence salience by considering the impact of the neighbor nodes based on the spreading activation model [12]. In [16], the iSpreadRank centrality demonstrates its performance superiority over other four studied centrality models.

In this work, we also apply iSpreadRank to calculate the sentence salience scores. There are three stages in the iSpreadRank computation: (1) initialization; (2) inference; and (3) prediction. In the initialization stage, iSpreadRank prepares a topic-similarity matrix A for the topic-similarity graph G according to the similarity measure. In A,

$$a_{i,j} = a_{j,i} = \begin{cases} 0 & \text{if } i = j \\ \mathrm{sim}(s_i, s_j) & \text{if } i \neq j \end{cases}. \tag{13}$$

Then A is transformed to a stochastic matrix R in which

$$r_{i,j} = \frac{a_{i,j}}{\sum_k a_{i,k}}. \tag{14}$$

Since all isolated node have been removed from G, $\sum_j r_{i,j} = 1$.

In the inference stage, the matrix R is used to calculate the spreading influences of the neighbor nodes. The inference is an iterative process to update the activation status of nodes by considering the spreading influences. Let V^t in G represent the activation status of nodes at iteration t and V^0 be the initial activation, V^t is calculated as:

$$V^t = V^0 + MV^{t-1}, M = \sigma R^T, \tag{15}$$

where σ ($0 \leq \sigma < 1$) is the decay factor to determine the propagation efficiency. It is assigned to 0.7 as [12]. The elements of V^0 are all initialized as 1 [16]. The termination condition of the iteration is reached when

$$\sum_i |V_i^t - V_i^{t-1}| \leq \varepsilon, \tag{16}$$

where ε=0.0001 is the threshold to control the termination condition. Finally, iSpreadRank centrality scores are all decided according to the iterative computation of Eqs. (15) and (16). In the prediction stage, the scores are then ranked.

3.5 Summary Generation

Two issues need to be concerned for summary generation. The first is the size of the summary. This is decided according to the compression rate *R*, which is the ratio of the summary length over the source length. The second is how to avoid that the summary includes sentences having redundant information. CSIS [9] is used for these two issues.

Fig. 3 shows the CSIS algorithm. Each sentence has a corresponding topic-salience score in each network model. The sentence s_c with the top score is the candidate sentence. In CSIS, a similarity threshold C_R is used to decide whether s_c is semantically redundant to any sentence s_j selected in the summary. With CSIS, all semantically redundant candidate sentences will be discarded. In this work, C_R is 0.7 as the previous studies [9, 16].

```
Input: a rank list L of sentences
Output: the summary Sum
Initialize:
    set Sum = ∅
Summarize:
    while the required compression rate R is not met
        s_c ← the candidate sentence having the highest score in L
        if max_{s_j ∈ Sum} sim(s_c, s_j) < C_R
            add s_i to Sum
        else
            omit s_i
        endif
        remove s_i from L
output the summary Sum
```

Fig. 3. The CSIS algorithm for summary generation.

4. EXPERIMENTS

To evaluate the performance of the proposed summarization approach, we have conducted empirical experiments on four official DUC (Document Understanding Conference) datasets. We discuss two issues in the experiments. First, we investigate the influence of different configurations of the proposed approach. Second, we explore the effectiveness of the proposed approach by comparing it with previous work.

4.1 Datasets and Evaluation Metrics

We used the official 2004-2007 DUC datasets in the experiments. Table 1 shows the details of the datasets.

Table 1. The details of the experimental datasets.

	DUC 2004	DUC 2005	DUC 2006	DUC 2007
# of collections	50	50	50	45
# of document/collection	10	25-50	25	25
Summary length	665 bytes	250 words	250 words	250 words

For performance evaluation, this work uses the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) toolkit 1.5.5 [41, 42]. ROUGE has been used widely in many studies since its first use in DUC 2004. It measures recall-based scores using n-gram co-occurrence statistics between the generated summary and a set of reference summaries, such as the scores for the 1-gram, 2-gram, 3-gram, 4-gram, and longest common substring units. In this paper, the ROUGE-1 (unigram-based), ROUGE-2 (bi-gram-based), and ROUGE-SU4 (skip-bigrams of 4) performance of the proposed probabilistic topic-based network models are measured to show their characteristics. However, we mainly discuss the ROUGE-1 performance of the proposed approach and previous work, because the ROUGE-1 measure has been shown to have high correlation with human assessments in the past studies [41, 42].

The ROUGE toolkit has many parameters in performance evaluation. For example, we use the parameter settings "`-e data -c 95 -b 665 -x -m -n 1`" to calculate ROUGE-1 scores for DUC 2004, where "–b 665" indicates that the maximum length of the summary is 665 bytes, and "-m" specifies the usage of stemming. The parameters "`-e data -n 1 -x -m -u -c 95 -r 1000 -f A -p 0.5 -t 0`" are used to calculate ROUGE-1 scores for DUC 2005-2007. These settings follow the settings of DUC 2004-2007 competition requirements.

4.2 Results and Discussion

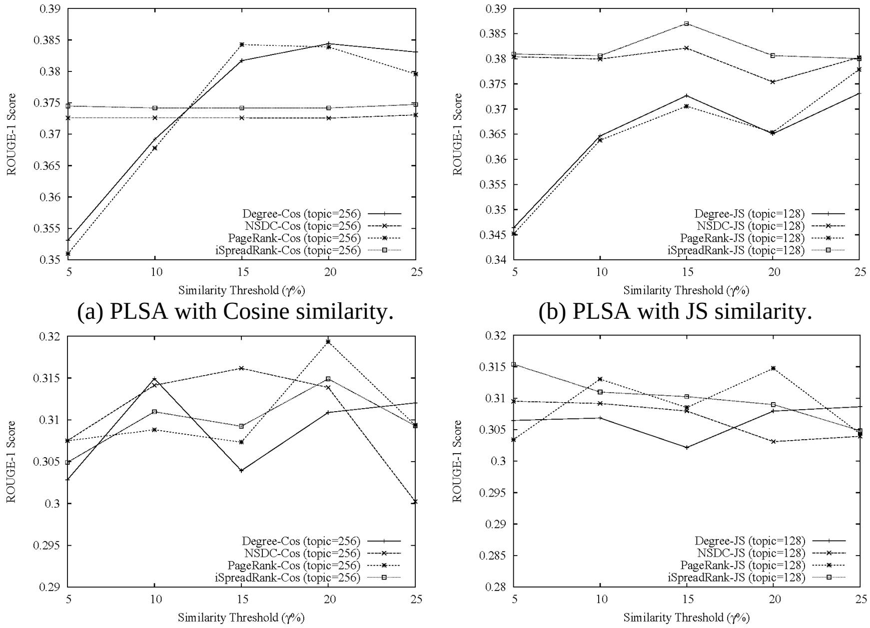
In the experiments, we have investigated two topic models, PLSA and LDA, combined with four network centrality models: Degree centrality (Degree), Normalized Similarity-based Degree centrality (NSDC), PageRank centrality (PageRank), and iSpreadRank centrality (iSpreadRank). In the PLSA implementation in Java using the EM algorithm with random initialization, we notice that the initialization influences PLSA. Therefore, we avoid this problem by averaging 5 random initializations as pointed in [29, 43]. For LDA, we use MALLET 2.0.7 with default hyper-parameters α=topic number/50 and β=0.01 [44].

We also implemented a baseline according to [16] in which each sentence is expressed as a TF-IDF vector and four centrality scores are calculated respectively. In the baseline model, only Cosine similarity is used with a similarity threshold S_t to decide the existence of links. If $sim(s_i,s_j) \geq S_t$, the edge e_{ij} is considered in the network-based centrality computation. The parameter configurations for all models are shown in Table 2. Most of the settings follow the previous studies [5, 9, 16, 29, 32].

Fig. 4 shows the ROUGE-1 scores of the 16 various configurations of the proposed approach for DUC 2004. Only DUC 2004 is presented to demonstrate the characteristics of different topic models and network models due to the length consideration.

Table 2. Parameter configurations in the experiments.

Topic Number	K=8,16,32,64,128,256 [29]
Bottom Topic Similarity Threshold	γ = 5%,10%,15%,20%,25%
Similarity Threshold in Baseline	S_t =0.1 [5]
Damping Factor	d = 0.15 [32]
Decay Factor	σ = 0.7 [16]
iSpreadRank Termination Control	ε = 0.001 [45]
CSIS Redundancy Threshold	C_R = 0.7 [9, 16]



(a) PLSA with Cosine similarity. (b) PLSA with JS similarity.

(c) LDA with Cosine similarity. (d) LDA with JS similarity.

Fig. 4. The ROUGE-1 recall scores for DUC 2004 using PLSA and LDA with Cosine and JS similarity.

Tables 3-6 show the best ROUGE-1 scores of the 16 various configurations of the proposed models for DUC 2004-2007 and the corresponding ROUGE-2 and ROUGE-SU4 scores. The highest scores are in a bold type.

Table 3. The best ROUGE-1, ROUGE-2, and ROUGE-SU4 scores of the proposed probabilistic topic-based network models and the baseline for DUC 2004.

		ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline (TFIDF+Network)	TFIDF-Degree-Cos	0.35792	0.08350	0.12419
	TFIDF-NSDC-Cos	0.35375	0.07224	0.11614
	TFIDF-PageRank-Cos	0.36575	0.08517	0.12569
	TFIDF-iSpreadRank-Cos	0.37028	0.09004	0.12966
PLSA+Network	PLSA-Degree-Cos	0.38440	0.09010	0.13395
	PLSA-Degree-JS	0.37621	0.08037	0.12752
	PLSA-NSDC-Cos	0.37878	0.08472	0.13024
	PLSA-NSDC-JS	0.38494	0.08768	0.13397
	PLSA-PageRank-Cos	0.38426	0.08853	0.13402
	PLSA-PageRank-JS	0.37832	0.07998	0.12840
PLSA+Network	PLSA-iSpreadRank-Cos	0.38087	0.08297	0.13020
	PLSA-iSpreadRank-JS	**0.38701**	**0.09277**	**0.13668**
LDA+Network	LDA-Degree-Cos	0.31493	0.04577	0.09246
	LDA-Degree-JS	0.30864	0.04409	0.09068

Table 3. (Cont'd) The best ROUGE-1, ROUGE-2, and ROUGE-SU4 scores of the proposed probabilistic topic-based network models and the baseline for DUC 2004.

	LDA-NSDC-Cos	0.31616	0.04786	0.09453
	LDA-NSDC-JS	0.31305	0.04536	0.09294
	LDA-PageRank-Cos	0.31933	0.05102	0.09615
	LDA-PageRank-JS	0.31474	0.04658	0.09336
	LDA-iSpreadRank-Cos	0.31490	0.04502	0.09351
	LDA-iSpreadRank-JS	0.31539	0.04519	0.09395

Table 4. The best ROUGE-1, ROUGE-2, and ROUGE-SU4 scores of the proposed probabilistic topic-based network models and the baseline for DUC 2005.

		ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline (TFIDF+Network)	TFIDF-Degree-Cos	0.36376	0.06735	0.12291
	TFIDF-NSDC-Cos	0.36293	0.06161	0.11849
	TFIDF-PageRank-Cos	0.36685	0.06707	0.12352
	TFIDF-iSpreadRank-Cos	0.36725	0.06824	0.12425
PLSA+Network	PLSA-Degree-Cos	0.38510	**0.07269**	0.13239
	PLSA-Degree-JS	0.38060	0.06700	0.12834
	PLSA-NSDC-Cos	0.37368	0.06276	0.12402
	PLSA-NSDC-JS	0.38576	0.07155	0.13279
	PLSA-PageRank-Cos	0.38456	0.07145	0.13221
	PLSA-PageRank-JS	0.38049	0.07217	0.13092
	PLSA-iSpreadRank-Cos	0.37473	0.06389	0.12393
	PLSA-iSpreadRank-JS	**0.38628**	0.07248	**0.13316**
	LDA-Degree-Cos	0.32134	0.03942	**0.09609**
	LDA-Degree-JS	0.31900	0.04140	**0.09718**
	LDA-NSDC-Cos	0.31873	0.04035	**0.09585**
	LDA-NSDC-JS	0.32380	0.04404	**0.09986**
	LDA-PageRank-Cos	0.32476	0.04144	**0.09832**
	LDA-PageRank-JS	0.32223	0.04134	**0.09789**
	LDA-iSpreadRank-Cos	0.31752	0.03826	**0.09618**
	LDA-iSpreadRank-JS	0.32130	0.04137	**0.09768**

Table 5. The best ROUGE-1, ROUGE-2, and ROUGE-SU4 scores of the proposed probabilistic topic-based network models and the baseline for DUC 2006.

		ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline (TFIDF+Network)	TFIDF-Degree-Cos	0.39414	0.08573	0.14106
	TFIDF-NSDC-Cos	0.39491	0.07935	0.13639
	TFIDF-PageRank-Cos	0.39934	0.08582	0.14207
	TFIDF-iSpreadRank-Cos	0.39749	0.08493	0.14054
PLSA+Network	PLSA-Degree-Cos	0.41315	0.08642	0.14696
	PLSA-Degree-JS	0.40721	0.08526	0.14323
	PLSA-NSDC-Cos	0.41296	0.08748	**0.14743**
	PLSA-NSDC-JS	0.41218	0.08740	0.14729

Table 5. (Cont'd) The best ROUGE-1, ROUGE-2, and ROUGE-SU4 scores of the proposed probabilistic topic-based network models and the baseline for DUC 2006.

	PLSA-PageRank-Cos	**0.41355**	0.08629	0.14703
	PLSA-PageRank-JS	0.40662	0.08437	0.14276
	PLSA-iSpreadRank-Cos	0.41172	**0.08795**	0.14708
	PLSA-iSpreadRank-JS	0.41143	0.08753	0.14740
LDA+Network	LDA-Degree-Cos	0.35658	0.05789	0.11399
	LDA-Degree-JS	0.35159	0.05497	0.10977
	LDA-NSDC-Cos	0.35204	0.05603	0.11152
	LDA-NSDC-JS	0.35232	0.05388	0.11080
LDA+Network	LDA-PageRank-Cos	0.35561	0.05780	0.11280
	LDA-PageRank-JS	0.35185	0.05441	0.11046
	LDA-iSpreadRank-Cos	0.35350	0.05494	0.11132
	LDA-iSpreadRank-JS	0.35181	0.05519	0.11063

Table 6. The best ROUGE-1, ROUGE-2, and ROUGE-SU4 scores of the proposed probabilistic topic-based network models and the baseline for DUC 2007.

		ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline (TFIDF+Network)	TFIDF-Degree-Cos	0.41382	0.09872	0.15507
	TFIDF-NSDC-Cos	0.41187	0.09092	0.14818
	TFIDF-PageRank-Cos	0.41560	0.09723	0.15387
	TFIDF-iSpreadRank-Cos	0.41419	0.09693	0.15350
	PLSA-Degree-Cos	0.42890	0.10150	0.15860
	PLSA-Degree-JS	0.42878	0.10375	0.16078
	PLSA-NSDC-Cos	0.42889	0.10153	0.15933
	PLSA-NSDC-JS	0.43313	0.10556	0.16345
	PLSA-PageRank-Cos	0.42800	0.10516	0.16069
	PLSA-PageRank-JS	0.43043	0.10499	0.16155
	PLSA-iSpreadRank-Cos	0.42793	0.10065	0.15916
	PLSA-iSpreadRank-JS	0.43024	0.10295	0.16164
	LDA-Degree-Cos	0.37304	0.06734	0.12415
	LDA-Degree-JS	0.36873	0.06559	0.12329
	LDA-NSDC-Cos	0.36831	0.06374	0.12244
	LDA-NSDC-JS	0.36931	0.06704	0.12272
	LDA-PageRank-Cos	0.36804	0.06132	0.12068
	LDA-PageRank-JS	0.36827	0.06621	0.12236
	LDA-iSpreadRank-Cos	0.36756	0.06450	0.12139
	LDA-iSpreadRank-JS	0.36726	0.06611	0.12204

From the results, we can have two observations. First, although PLSA may have overfitting problems, PLSA outperforms LDA in all configurations of the network models and the similarity models, and achieves the best ROUGE-1 performance for all DUC 2004-2007 datasets. As we manually investigate the topic distributions calculated by PLSA and LDA, we find that the distributions of the PLSA topics have significant variances. Many topics conveying salient information for summarization can be discriminated. In contrary, the distributions of the LDA topics do not have significant differences. Therefore, many topic-based vector nodes decided by LDA are close in the network models. The deficit of the topic discriminative capability of LDA thus makes the network models consider more insignificant sentences. The investigation shows that PLSA has better topic discriminative capability than LDA for the multi-document summarization task. Moreover, the performance of TFIDF is better than that of LDA because of the same situation. One possible reason for the poor performance of LDA may be because it severely suffers from the data sparsity problem existing in short text [46]. However, this problem is mitigated in PLSA for DUC datasets because PLSA may capture more details of the topic distributions due to its maximum-likelihood characteristics. A similar observation has been noticed for the short text problem [47]. In that work, PLSA achieves better performance than the simple bag-of-word model for short text when the number of the training documents is small.

Second, the performance also shows that these four network centrality models can be classified into two classes: the degree-based and the topic-similarity-based. The Degree and PageRank centrality models belong to the degree-based class, and NSDC and iSpreadRank are in the topic-similarity-based class. When PLSA is used as the topic model, NSDC and iSpreadRank have stable and close ROUGE-1 performance for various similarity thresholds; Degree and PageRank have very close but unstable performance. Although these network models have unstable ROUGE-1 performance in LDA, the models of the same class have similar performance patterns.

(1)Anwar, 51, was arrested Sept. 20 under the Internal Security Act, which allows indefinite detent
(2)Anwar was fired by Prime Minister Mahathir Mohamad on Sept. 2 after the two differed on eco
(3)Mahathir fired Anwar on Sept. 2 from his posts as deputy prime minister and finance minister, s
(4)On Sept. 2, Malaysia’s Prime Minister Mahathir Mohamad fired Anwar, calling him morally un
(5)“I told them that I don't have to see my husband.

(a)TFIDF-iSpreadRank-Cos.

(1)Anwar, 51, was arrested Sept. 20 under the Internal Security Act, which allows indefinite detent
(2)On Sept. 2, Malaysia's Prime Minister Mahathir Mohamad fired Anwar, calling him morally un
(3)Anwar was fired by Prime Minister Mahathir Mohamad on Sept. 2 after the two differed on eco
(4)Jailed, beaten and facing trial on 10 sexual misconduct and corruption charges, ousted Deputy P
(5)Anwar said police beat him in custody.

(b) PLSA-iSpreadRank-JS.

(1)At least two ASEAN leaders, Philippine President Joseph Estrada and Indonesian President B.J.
(2)Munawar Ahmad Aness, a friend and speech writer of Anwar Ibrahim, pleaded guilty to the cha
(3)Anwar, 51, was arrested Sept. 20 under the Internal Security Act, which allows indefinite detent
(4)Pillai, who runs a popular website on local politics, isn't surprised by the aggressive march towa
(5)Malaysian journalist M.G.G.

(c) LDA-iSpreadRank-Cos.

Fig. 5. The generated summary examples of three probabilistic topic-based network models for the article set `d30011t` in DUC 2004.

Fig. 5 shows the generated summary examples of three probabilistic topic-based network models for the article set `d30011t` in DUC 2004: TFIDF-iSpreadRank-Cos, PLSA-iSpreadRank-JS, and LDA-iSpreadRank-Cos. The order of the sentences in each generated summary is ranked according to their topic-salience scores in CSIS. The results show that TFIDF-iSpreadRank-Cos and LDA-iSpreadRank-Cos select some sentences that convey less information about the news, such as the fifth sentence in LDA-iSpreadRank-Cos. In contrary, the sentences generated by PLSA-iSpreadRank-JS are more pertinent to the news.

In order to illustrate the effectiveness of the proposed probabilistic topic-based network models, Table 7 shows the ROUGE-1 performance comparison of the proposed models with previous summarization schemes. In the table, the configuration PLSA-iSpreadRank-JS is used for performance comparison on all datasets because it achieves the best performance in DUC 2004 and 2005. Moreover, the best PLSA configurations for DUC 2006 and 2007, PLSA-PageRank-Cos and PLSA-NSDC-JS, are also included for comparison.

Table 7. Performance comparison of the proposed probabilistic topic-based network models with previous schemes for DUC 2004-2007.

Dataset	Systems	ROUGE-1
DUC 2004	Best Machine in DUC 2004 (SID=65)	0.38224 (5)
	Runner-up Machine in DUC 2004 (SID=104)	0.37443 (8)
	Third-place Machine in DUC 2004 (SID=35)	0.37430 (9)
	Top score of LexRank [5]	0.3830 (4)
	Top score of iSpreadRank [12]	0.38068 (6)
DUC 2004	Bi-PLSAS [48]	**0.38853** (1)
	Cai & Li [49]	0.37475 (7)
	Pos+iSpreadRank [16]	0.38634 (3)
	PLSA-iSpreadRank-JS	0.38701 (2)
DUC 2005	Best Machine in DUC 2005 (SID=15)	0.38036 (5)
	Runner-up Machine in DUC 2005 (SID=4)	0.37910 (6)
	Third-place Machine in DUC 2005 (SID=17)	0.37362 (7)
	Bi-PLSAS [48]	0.36028 (9)
	TopicAffinityRank1 [11]	0.38354 (4)
	DESAMC+DocSum [13]	0.3937 (2)
	Cai & Li [49]	0.36451 (8)
	MA-MultiSumm [50]	**0.4001** (1)
	PLSA-iSpreadRank-JS	0.38628 (3)
DUC 2006	Best Machine in DUC 2006 (SID=24)	0.40980 (5)
	Runner-up Machine in DUC 2006 (SID=12)	0.40488 (7)
	Third-place Machine in DUC 2006 (SID=23)	0.40440 (8)
	Bi-PLSAS [48]	0.39384 (9)
	DESAMC+DocSum [13]	**0.4345** (1)
	Cai & Li [49]	0.40581 (6)
	MA-MultiSumm [50]	0.4195 (2)
	PLSA-PageRank-Cos	0.41355 (3)
	PLSA-iSpreadRank-JS	0.41143 (4)
DUC 2007	Best Machine in DUC 2007 (SID=24)	**0.45258** (1)
	Runner-up Machine in DUC 2007 (SID=15)	0.44508 (2)
	Third-place Machine in DUC 2007 (SID=4)	0.43417 (3)
	Cai & Li [49]	0.41622 (7)

Table 7. (Cont'd) Performance comparison of the proposed probabilistic topic-based network models with previous schemes for DUC 2004-2007.

	CTSUM [39]	0.43101 (5)
	Hybrid-TM [31]	0.381 (8)
	PLSA-NSDC-JS	0.43313 (4)
	PLSA-iSpreadRank-JS	0.43024 (6)

Table 7 presents the top 3 participating systems for DUC 2004-2007, in which *SID* is the peer code numbers of the participants. The bold numbers show the highest ROUGE-1 scores of these systems. The ROUGE-1 data of the compared schemes are obtained directly from the corresponding reports or papers. The number between parentheses is the ranking of each scheme in Table 7.

Because the ROUGE-1 performance has been shown to have high correlation with human assessments than other ROUGE metrics in the past studies [41, 42], this paper compares the ROUGE-1 performance of the proposed probabilistic topic-based network models with previous approaches for DUC 2004-2007. As shown in Table 7, the proposed probabilistic topic-based network models consistently achieve high ROUGE-1 performance for all datasets. Moreover, the proposed probabilistic topic-based network models outperform the best DUC-participating systems for DUC 2004-2006. Although Bi-PLSAS has the top performance for DUC 2004, the proposed approach outperforms Bi-PLSAS in DUC 2005-2006. Although the proposed approach takes the third place in DUC 2005-2006, both DESAMC+DocSum and MA-MultiSumm are two evolutionary-based optimization schemes which need a large number of evaluations of the objective functions or complicated parameter tuning. For DUC 2007, the proposed approach outperforms other recently devised summarization schemes.

5. CONCLUSIONS

Automatic multi-document summarization is a challenging problem that has otained significant attention in the research domain of text summarization. Probabilistic topic models and network models have demonstrated their effectiveness in multi-document summarization [5, 12, 14-16, 18, 19, 30]. However, only few studies discuss the integration of two models [14, 15], and they all consider only LDA with two popular network models, PageRank and HITS.

This paper proposes an extractive approach considering both probabilistic topic models and network models to generate the summary. Two probabilistic topic models and four network models are investigated. Comprehensive experimental studies are conducted with the DUC 2004-2007 datasets. The experimental results show that the PLSA-based network approach outperforms the TF-IDF baseline approach on all datasets. A systematic comparison between two representative topic models (PLSA and LDA) is also conducted. The results show that PLSA outperforms LDA by effectively identifying crucial topics for the datasets. Compared with other state-of-the-art approaches on all datasets, the PLSA-based network approach can stably have high performance.

In our future work, more experiments will be conducted on other datasets to validate the generalization of the proposed probabilistic topic-based network approach. Enhancements based on the proposed approach will be also investigated.

ACKNOWLEDGEMENT

The authors would like to express their sincere thanks to anonymous reviewers, Prof. Hung-Yu Kao, and Prof. Wen-Hsiang Lu for their precious comments and suggestions.

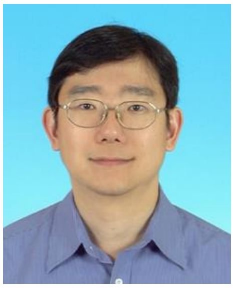
REFERENCES

1. U. Hahn and I. Mani, "The challenges of automatic summarization," *Computer*, Vol. 33, 2000, pp. 29-36.
2. E. Lloret and M. Palomar, "Text summarisation in progress: A literature review," *Artificial Intelligence Review*, Vol. 37, 2012, pp. 1-41.
3. A. Nenkova and K. McKeown, "A survey of text summarization techniques," in C. C. Aggarwal and C.X. Zhai, (ed.), *Mining Text Data*, Springer, NY, 2012, pp. 43-76.
4. S. Sekine and C. Nobata, "A Survey for multi-document summarization," in *Proceedings of the HLT-NAACL 03 on Text Summarization Workshop*, Vol. 5, 2003, pp. 65-72.
5. G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, Vol. 22, 2004, pp. 457-479.
6. J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, "Multi-document summarization by sentence extraction," in *Proceedings of NAACL-ANLP Workshop on Automatic Summarization*, 2000, pp. 40-48.
7. K. McKeown and D. R. Radev, "Generating summaries of multiple news articles," in *Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 74-82.
8. D. R. Radev and K. R. McKeown, "Generating natural language summaries from multiple on-line sources," *Computational Linguistics*, Vol. 24, 1998, pp. 470-500.
9. D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, Vol. 40, 2004, pp. 919-938.
10. I. Mani and E. Bloedorn, "Summarizing similarities and differences among related documents," *Information Retrieval*, Vol. 1, 1999, pp. 35-67.
11. X. Wan and J. Xiao, "Towards a unified approach based on affinity graph to various multi-document summarizations," in *Proceedings of the 11th European Conference on Digital Libraries*, 2007, pp. 297-308.
12. J.-Y. Yeh, H.-R. Ke, and W.-P. Yang, "iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network," *Expert Systems with Applications*, Vol. 35, 2008, pp. 1451-1462.
13. R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization," *Knowledge-Based Systems*, Vol. 36, 2012, pp. 21-38.
14. D. Gao, W. Li, O. You, and R. Zhang, "LDA-based topic formation and topic-sentence reinforcement for graph-based multi-document summarization," in *Proceedings of the 8th Asia Information Retrieval Societies Conference*, 2012, pp. 376-385.
15. Y. Pei, W. Yin, and L. Huang, "Generic multi-document summarization using topic-oriented information," in *Proceedings of the 12th Pacific Rim International Conference on Trends in Artificial Intelligence*, 2012, pp. 435-446.
16. J.-Y. Yeh, W.-P. Yang, H.-R. Ke, and P.-C. Cheng, "Extraction-based news summarization using sentence centrality in the sentence similarity network," *Journal of Information Management*, Vol. 21, 2014, pp. 271-304.

17. D. M. Blei, A. Y. Ng, and M. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.
18. R. Arora and B. Ravindran, "Latent Dirichlet allocation based multi-document summarization," in *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, 2008, pp. 91-97.
19. A. Celikyilmaz and D. Hakkani-Tur, "A hybrid hierarchical model for multi-document summarization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 815-824.
20. Z. Liu, W. Huang, Y. Zheng, and M. Sun, "Automatic keyphrase extraction via topic decomposition," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 366-376.
21. T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50-57.
22. L. Hennig, "Content modeling for automatic document summarization," Ph.D. Dissertation, Department of Technische Universität Berlin, 2011.
23. Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA," *Information Retrieval*, Vol. 14, 2011, pp. 178-203.
24. T. Masada, S. Kiyasu, and S. Miyahara, "Comparing LDA with pLSI as a dimensionality reduction method in document clustering," in *Proceedings of the 3rd International Conference on Large-Scale Knowledge Resources*, 2008, pp. 13-26.
25. J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary, "Topic-focused multi-document summarization using an approximate oracle score," in *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, 2006, pp. 152-159.
26. S. Harabagiu and F. Lacatusu, "Using topic themes for multi-document summarization," *ACM Transactions on Information Systems*, Vol. 28, 2010, pp. 1-47.
27. S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, Vol. 41, 1990, pp. 391-407.
28. D. Hull, "Improving text retrieval for the routing problem using latent semantic indexing," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 282-291.
29. L. Hennig, "Topic-based multi-document summarization with probabilistic latent semantic analysis," in *Proceedings of International Conference on Recent Advances in Natural Language Processing*, 2009, pp. 144-149.
30. A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Hidden topic Markov models," in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, 2007, pp. 163-170.
31. J.-A. Xu, J.-M. Liu, and K. Araki, "A hybrid topic model for multi-document summarization," *IEICE Transactions on Information and Systems*, Vol. E98-D, 2015, pp. 1089-1094.
32. S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, Vol. 30, 1998, pp. 107-117.
33. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, Vol. 46, 1999, pp. 604-632.

34. R. Mihalcea and P. Tarau, “A language independent algorithm for single and multiple document summarization,” in *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, 2005, pp. 19-24.
35. J. R. Anderson, “A spreading activation theory of memory,” *Journal of Verbal Learning and Verbal Behavior*, Vol. 22, 1983, pp. 261-295.
36. X. Wan and J. Yang, “Multi-document summarization using cluster-based link analysis,” in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 299-306.
37. T.-Y. Liu and W.-Y. Ma, “Webpage importance analysis using conditional Markov random walk,” in *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 2005, pp. 515-521.
38. F. Fukumoto, A. Sakai, and Y. Suzuki, “Eliminating redundancy by spectral relaxation for multi-document summarization,” in *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, 2010, pp. 98-102.
39. X. Wan and J. Zhang, “CTSUM: Extracting More Certain Summaries for News Articles,” in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014, pp. 787-796.
40. T. L. Griffiths and M. Steyvers, “Finding scientific topics,” in *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, 2004, pp. 5228-5235.
41. C.-Y. Lin and E. Hovy, “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*, Vol. 1, 2003, pp. 71-78.
42. C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74-81.
43. T. Brants, F. Chen, and I. Tsochantaridis, “Topic-based Document Segmentation with Probabilistic Latent Semantic Analysis,” in *Proceedings of the 11th International Conference on Information and Knowledge Management*, 2002, pp. 211-218.
44. A. K. McCallum, “MALLET: A Machine Learning for Language Toolkit,” http://mallet.cs.umass.edu, 2002.
45. J.-Y. Yeh, W.-P. Yang, H.-R. Ke, P.-C. Cheng, and C.-H. Yu, “News summarization based on graphical network models,” in *Proceedings of Annual Conference on Practical Information Management*, 2013, pp. 323-340.
46. L. Hong and B. D. Davison, “Empirical Study of Topic Modeling in Twitter,” in *Proceedings of the 1st Workshop on Social Media Analytics*, 2010, pp. 80-88.
47. Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. Chen, “Toward multi-modal music emotion classification,” in *Proceedings of the 9th Pacific Rim Conference on Multimedia*, 2008, pp. 70-79.
48. C. Shen, T. Li, and C. H. Q. Ding, “Integrating clustering and multi-document summarization by bi-mixture Probabilistic Latent Semantic Analysis (PLSA) with sentence bases,” in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011, pp. 914-920.
49. X. Cai and W. Li, “Ranking through clustering: an integrated approach to multi-document summarization,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, 2013, pp. 1424-1433.

50. M. Mendoza, C. Cobos, E. León, M. Lozano, F. Rodríguez, and E. Herrera-Viedma, "A new memetic algorithm for multi-document summarization based on CHC algorithm and greedy search," in *Proceedings of the 13th Mexican International Conference on Artificial Intelligence*, 2014, pp. 125-138.



Cheng-Zen Yang (楊正仁) received the B.S. and M.S. degrees from Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, in 1988 and 1990, respectively. Then he received his Ph.D. degree from Department of Computer Science and Information Engineering, National Taiwan University in 1996. Since 1998, he has been an Assistant Professor at Yuan Ze University. His research interests include web technology, text mining, software engineering, and high-speed computing. He is a member of ACM and IEEE.

Jhih-Sheng Fan (范植昇) received the B.S. degree from Department of Computer Science and Engineering, Yuan Ze University in 2015. Currently, he is a graduate student at Department of Computer Science and Information Engineering, National Cheng Kung University. Since 2013, he has joined research projects on text mining and natural language processing. His research interests include text mining, web mining, and natural languange processing.

Yu-Fan Liu (劉育凡) received the B.S. degree from Department of Computer Science and Engineering, Yuan Ze University in 2015. Currently, she is a graduate student at Department of Computer Science, National Chiao Tung University. Her research interests include text mining, embedded systems, and mobile networking.