SentTopic-MultiRank: a novel ranking model for multi-document summarization

Wenpeng Yin Yulong Pei Fan Zhang Lian′en Huang
The Shenzhen Key Lab for Cloud Computing Technology & Applications (SPCCTA)
Shenzhen Graduate School
Peking University, Shenzhen 518055, P.R. China
{mr.yinwenpeng,paul.yulong.pei,fan.zhgf}@gmail.com, hle@net.pku.edu.cn

ABSTRACT
Extractive multi-document summarization is mostly treated as a sentence ranking problem. Existing graph-based ranking methods for key-sentence extraction usually attempt to compute a global importance score for each sentence under a single relation. Motivated by the fact that both documents and sentences can be presented by a mixture of semantic topics detected by Latent Dirichlet Allocation (LDA), we propose SentTopic-MultiRank, a novel ranking model for multi-document summarization. It assumes various topics to be heterogeneous relations, then treats sentence connections in multiple topics as a heterogeneous network, where sentences and topics/relations are effectively linked together. Next, the iterative algorithm of MultiRank is carried out to determine the importance of sentences and topics simultaneously. Experimental results demonstrate the effectiveness of our model in promoting the performance of both generic and query-biased multi-document summarization tasks.

KEYWORDS: Multi-document summarization, Topic decomposition, Heterogeneous network.

Proceedings of COLING 2012: Technical Papers, pages 2977–2992,
COLING 2012, Mumbai, December 2012.

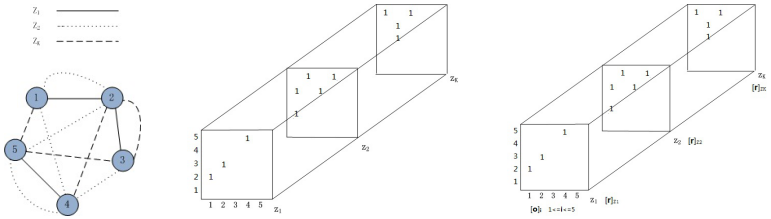
1 Introduction

Multi-document summarization (MDS), having been intensively studied in past decades, is the process of automatically creating a compressed version of a given document collection that provides useful information for a user. It could be mainly classified into two categories: generic MDS and query-focused MDS. The goal of generic MDS is to produce a summary of multiple documents about the same but unspecified topic, while query-focused task requires the generated summary could not only convey the main content of target corpus, but also bias to the information needs of a specific query/topic. Commonly, both of them are treated as a sentence ranking problem.

In most existing summarization systems, sentence ranking was conducted under a single relation assumption. Taking for instance the famous LexPageRank proposed in (Erkan and Radev, 2004). In that model, a sentence connectivity matrix was constructed based on cosine similarity. Then PageRank (Page et al., 1999) algorithm was directly applied to the cosine similarity graph to find the most prestigious sentences in a document. Another compelling example is (Wan et al., 2007), where a weighted network was formed on the sentences by using standard cosine similarity, then authors utilized manifold-ranking algorithm (Zhou et al., 2004) to spread ranking scores from a query to remaining sentences. Apparently, both above literatures were based on the assumption that all sentences existed under a unified relation.

Inspired by some work involving topic decomposition as well as multi-relational data where objects have interactions with others based on different relations, we attempt to map sentence relatedness within multiple topics to heterogeneous relations in this work. More specifically, we assume each topic as a single relation type, and construct an intra-topic sentence network for each relation type. There are many Information Retrieval and Data Mining tasks addressing multi-relational data. For instance, scholars cite other scholars in various conferences, and based on different academic fields, publications cite other publications on the basis of content analysis such as authorship, title, abstract and keywords, web pages link to each other via different anchor texts (Kolda and Bader, 2006). Such a complicated link structure can provide a way of incorporating multiple relations among objects into the derivation of object prestige or popularity. In Figure 1(a), we show an example of multi-relational sentence representation based on our SentTopic-MultiRank model. There are five sentences and K relations among them (K denotes the topic number in LDA, and we only provide three topics z_1, z_2 and z_K as an illustration). We can also represent such multi-relational objects in a tensor shape which is a multi-dimensional array. In the Figure 1(b), a three-way array is provided, where each two-dimensional plane represents an adjacency matrix for one type of relation. The network can be depicted as a tensor of size $5 \times 5 \times K$ where (i, j, k) entry is nonzero if the i^{th} sentence is related to the j^{th} sentence under k^{th} relation.

Then, how to determine the importance of those sentences and relations/topics? (Ng et al., 2011) proposed a general framework named MultiRank to deal with multi-relational data or the corresponding tensor representation for co-ranking purpose. According to that proposal, the MultiRank value of a sentence relies on the number and MultiRank values of all sentences that have multiple relations to this sentence, as well as the MultiRank values of those mutual relations. A sentence, connected via high MultiRank relations by sentences with high MultiRanks, receives a high MultiRank itself. Similarly, the MultiRank of a relation is dependent on which sentences to be linked and their MultiRank scores. A relation, connecting sentences with high MultiRanks, receives a high MultiRank itself. Similar to PageRank, MultiRank's idea



(a) An example of multi-relational sentences in a graph representation

(b) The corresponding tensor representation

(c) The tensor representation with prior

Figure 1: Illustration of SentTopic-MultiRank

is to imagine infinite random surfers in a multi-relational network, and derive a stationary probability distribution of objects and relations as evaluation scores for objects and relations, respectively. We integrate topic decomposition via LDA with MultiRank to produce a novel framework named SentTopic-MultiRank specifically for multi-document summarization.

For generic multi-document summarization, we directly make use of MultiRank algorithm to determine the ranking scores of sentences and topics/relations. While in query-oriented MDS, query bias must be considered. We creatively integrate LDA with generation probability between sentences to design query-oriented prior distributions for sentences and topics, respectively. Motivated by topic-sensitive PageRank (Haveliwala, 2003), as well as HAR (Xutao et al., 2012) which modified MultiRank based on HITS (Kleinberg, 1999) principle for query search, we embed acquired prior distributions into the SentTopic-MultiRank iterative process so that the final ranking results are more desirable for the query statement. Figure 1(c) gives an example for tensor representation of SentTopic-MultiRank with prior knowledge. We apply our approach to generic and query-biased multi-document summarization on standard datasets, experimental results show its good performance in generating high-quality summaries.

The rest of this paper is organized as follows: Section 2 introduces related work. Details of constructing and applying SentTopic-MultiRank model in multi-document summarization task are presented in Section 3. Section 4 gives the process of selecting sentences to generate summary. Finally, experiments and results are showed in Section 5.

2 Related work

In this section, we firstly introduce some representative work about multi-document summarization. Then, some literatures, relevant to topic decomposition, are also presented.

The centroid-based method (Radev et al., 2004) is one of the most popular extractive summarization methods. MEAD is an implementation of the centroid-based method that scores sentences based on features such as cluster centroids, position and TF-IDF. NeATS (Lin and Hovy, 2002) used sentence position, term frequency, topic signature and term clustering to select important content, and used MMR (Goldstein et al., 1999) to remove redundancy. Recently, graph-based methods have been proposed to rank sentences. LexPageRank (Erkan and Radev, 2004) and (Mihalcea and Tarau, 2005) are two such systems using algorithms similar to PageRank and HITS (Kleinberg, 1999) to compute sentence importance. With respect to query-

focused summarization, in (Saggion et al., 2003), a simple query-based scorer by computing the similarity value between each sentence and the query was incorporated into a generic summarizer to produce the query-based summary. (Erkan, 2006) came up with biased LexRank by incorporating prior knowledge into random walk for query summarization task. (Wan et al., 2007) exploited manifold-ranking process proposed in (Zhou et al., 2004) to implement the propagation of ranking scores from the query to remaining sentences, based on a weighted sentence connection network. The spread process was repeated until a global stable state was achieved, and all sentences obtained their final ranking scores. Recently, unsupervised deep learning was first investigated in (Liu et al., 2012) to deal with query summarization.

Furthermore, supervised learning approaches have also been successfully applied in document summarization, where the training data is available or easy to build. The most straightforward way is to regard the sentence extraction task as a binary classification problem. (Kupiec et al., 1995) developed a trainable summarization system which adopted various features and used a Bayesian classifier to learn the feature weights. The system performed better than other systems using only a single feature. (Zhou and Hovy, 2003) applied a HMM-based model and (Shen et al., 2007) proposed a conditional random field based framework. (Ouyang et al., 2007) designed methods for constructing training data based on human summaries and training sentence scoring models based on support vector regression (SVR). (Shen and Li, 2011) introduced a cost sensitive loss to improve ranking SVM, a type of learning-to-rank method, for extractive query-focused multi-document summarization.

Especially, we are mainly inspired by following pioneering work. In recent years, two algorithms were proposed to rank web pages by incorporating topic information within PageRank (Haveliwala, 2003; Nie et al., 2006). The method in (Haveliwala, 2003) decomposed PageRank into various topics, setting the preference values using some conditional probabilities. (Nie et al., 2006) proposed a more complicated ranking framework, where topical PageRanks were performed together. The rationale of (Nie et al., 2006) was, when surfing following a graph link from vertex w_i to w_j, the ranking score on topic z of w_i would have a higher probability to pass to the same topic of w_j and have a lower probability to pass to a different topic of w_j. When the inter-topic jump probability was 0, this method was identical to the approach in (Haveliwala, 2003). (Liu et al., 2010) explicitly defined a new graph-based framework, Topical PageRank, for the task of keyphrase extraction. It first ranked all key-phrases within each latent topic based on a topic-specific phrase graph, then re-ranked them by integrating corpus-topic distribution with intra-topic phrase ranking. Additionally, ToPageRank, a model similar with that of (Liu et al., 2010), was proposed in (Pei et al., 2012) specifically for summarization task. Indeed, a similarity between these literatures with our current idea lies in that we all consider to decompose entire corpus into multiple topics. Whereas, we do not aim to *break up the whole into parts* simply. In their work, there was usually no relation exerted to link each part, so the operation in each part was conducted independently and those objects would not interact with various kinds of relations. Hence, in essence, they still coped with ranking problem under a single relation presupposition, except that having narrowed the scope from a corpus to an individual topic. Differently, in SentTopic-MultiRank, not only are sentences linked with each other by weighted edges, all intra-topic sentence networks are also connected together, though they are considered to be under different relation types, so that we are able to identify the importance of sentences and topics/relations simultaneously. Further, we make full use of sentence-sentence relatedness, sentence-topic interaction and inter-topic impacts to improve the overall ranking performance.

3 SentTopic-MultiRank model

3.1 Topic decomposition of corpus via Latent Dirichlet Allocation (LDA)

In our work, topic model LDA (Blei et al., 2003) is utilized to represent document collection with a mixture of semantic topics. In LDA, it is assumed that observed words in each document are generated by a document-specific mixture of corpus-wide latent topics. We define our corpus of length W with the flat word vector $\mathbf{w} = w_1, \cdots, w_W$. At corpus position i, the element d_i in $\mathbf{d} = d_1, \cdots, d_W$ designates the document containing observed word w_i. Similarly, the vector $\mathbf{z} = z_1, \cdots, z_W$ defines the hidden topic assignments of each observed word. The number of latent topics is fixed to some K, and each topic $z = 1, \cdots, K$ is associated with a topic-word multinomial ϕ_z over the W-word vocabulary. Each ϕ multinomial is generated by a conjugate Dirichlet prior with parameter β. Each document $j = 1, \cdots, D$ is associated with a multinomial θ_j over K topics, which is also generated by a conjugate Dirichlet prior with parameter α. The full generative model is then given by

$$P(\mathbf{w}, \mathbf{z}, \phi, \theta | \alpha, \beta, \mathbf{d}) \propto \left(\prod_{z=1}^{K} p(\phi_z|\beta)\right)\left(\prod_{j=1}^{D} p(\theta_j|\alpha)\right)\left(\prod_{i=1}^{W} \phi_{z_i}(w_i)\theta_{d_i}(z_i)\right)$$

where $\phi_{z_i}(w_i)$ is the w_i-th element in vector ϕ_{z_i}, and $\theta_{d_i}(z_i)$ is the z_i-th element in vector θ_{d_i}. Given an observed corpus $(\mathbf{w}, \mathbf{d})$ and model hyperparameters (α, β), the typical modeling goal is to infer the latent variables $(\mathbf{z}, \phi, \theta)$.

While exact LDA inference is intractable, a variety of approximate schemes have been developed. In this work, we use GibbsLDA++[1], a C/C++ implementation of LDA using Gibbs Sampling, to detect latent topics. This sampling approach iteratively re-samples a new value for each latent topic assignment z_i, conditioned on the current values of all other $\mathbf{z}$ values. After a fixed number of iterations, we estimate the topic-word multinomials ϕ and the document-topic mixture weights θ from the final $\mathbf{z}$ sample, using the means of their posteriors given by

$$\phi_z(w) \propto n_{zw} + \beta$$
$$\theta_j(z) \propto n_{jz} + \alpha$$

where n_{zw} is the number of times word w is assigned to topic z, and n_{jz} is the number of times topic z is used in document j, with both counts being taken with respect to the final sample $\mathbf{z}$. The topic-word multinomials ϕ_z for each topic z are our learned topics; each document-topic multinomial θ_j represents the prevalence of topics within document j.

In experiments, we set $\alpha = 1$ and $\beta = 0.01$. There are already some literatures to study their impacts on LDA performance. Since we pay main attentions to SentTopic-MultiRank model, the influences of LDA hyperparameters α and β are not investigated any more.

3.2 Construction of SentTopic-MultiRank graph

In Section 3.1, we have detected various relation types/topics. The thing left to do in constructing SentTopic-MultiRank is to produce a sentence graph under each relation. Naturally, the sentence similarity should be relation-specific. To capture the similarities of two sentences

[1] GibbsLDA++: http://gibbslda.sourceforge.net

(e.g., x and y) on various latent topics, we represent each sentence as a probability distribution at each topic z ($1 \le z \le K$). Hence, we sample sparse unigram distributions from each ϕ_z using the words in x and y. Their probability distributions given topic-word distribution are denoted as $P_z^x = p(\mathbf{w}_x|z, \phi_z)$ with the word set $\mathbf{w}_x = (w_1, \cdots, w_{|x|})$ in x, and $P_z^y = p(\mathbf{w}_y|z, \phi_z)$ with the word set $\mathbf{w}_y = (w_1, \cdots, w_{|y|})$ in y.

The probability distributions per topic are constructed with only the words in x and y, and the probabilities of remaining words in vocabulary W are set to 0. The W dimensional word probabilities are the expected posteriors obtained from LDA model. Hence, $p_z^x = (\phi_z(w_1), \cdots, \phi_z(w_{|x|}), 0, 0, \cdots) \in (0,1)^W$, $p_z^y = (\phi_z(w_1), \cdots, \phi_z(w_{|y|}), 0, 0, \cdots) \in (0,1)^W$. Given a topic z, the similarity between p_z^x and p_z^y is measured via transformed radius (TR). We first measure the divergence at each topic using TR based on Kullback-Liebler (KL) divergence:

$$TR(p_z^x, p_z^y) = KL(p_z^x || \frac{p_z^x + p_z^y}{2}) + KL(p_z^y || \frac{p_z^x + p_z^y}{2}) \quad (1)$$

where $KL(m||n) = \sum_i m_i \log \frac{m_i}{n_i}$. Then TR is transformed into similarity (Manning et al., 1999):

$$Sim(p_z^x, p_z^y) = 10^{-TR(p_z^x, p_z^y)} \quad (2)$$

Here, we adopt TR rather than the commonly used KL for the reason that with TR there is no problem with infinite values since $\frac{p_z^x + p_z^y}{2} \neq 0$ if either $p_z^x \neq 0$ or $p_z^y \neq 0$, and it is also symmetric, i.e., $TR(x, y) = TR(y, x)$. With the gotten sentence similarities under various topics, we could construct graphs like Figures 1(a) and 1(b) to demonstrate our SentTopic-MultiRank model.

3.3 Sentence ranking using MultiRank algorithm

There are many Data Mining and Machine Learning issues in multi-relational data where objects interact with others under different relations. The work in (Ng et al., 2011) proposed a framework, MultiRank, to determine the importance of both objects and relations simultaneously based on a probability distribution computed from multi-relational data. In our multi-relational sentence network, we apply MultiRank algorithm to derive the importance of sentences and topics. As we analyze sentences under multiple relations and also consider interaction between relations based on sentences, we make use of rectangular tensors to represent them as Figure 1(b). First, we introduce the MultiRank iterative algorithm.

Let R be the real field. We call $\mathscr{A} = (a_{s_1,s_2,z_1})$ where $a_{s_1,s_2,z_1} \in R$, for $s_i = 1, \cdots, n$, $i = 1, 2$ and $z_1 = 1, \cdots, K$, a real (2,1)th order $(n \times K)$-dimensional rectangular tensor. In this setting, we refer (s_1, s_2) to the indices for sentences/objects and z_1 to be the index for topic/relation. For instance, five sentences ($n = 5$) and three relations ($K = 3$) are demonstrated in Figure 1.

First, two transition probability tensors $\mathscr{O} = (o_{s_1,s_2,z_1})$ and $\mathscr{R} = (r_{s_1,s_2,z_1})$ are constructed with respect to sentences and topics by normalizing the entries of $\mathscr{A}$ as follows:

$$o_{s_1,s_2,z_1} = \frac{a_{s_1,s_2,z_1}}{\sum_{i=1}^{m} a_{i,s_2,z_1}}, \quad s_1 = 1, 2, \cdots, n$$

$$r_{s_1,s_2,z_1} = \frac{a_{s_1,s_2,z_1}}{\sum_{j=1}^{K} a_{s_1,s_2,j}}, \quad z_1 = 1, 2, \cdots, K$$

Then we derive the following probabilities:

$$\text{Prob}[X_t = s_1] = \sum_{s_2=1}^{n} \sum_{z_1=1}^{K} o_{s_1,s_2,z_1} \times \text{Prob}[X_{t-1} = s_2, Y_t = z_1] \tag{3}$$

$$\text{Prob}[Y_t = z_1] = \sum_{s_1=1}^{n} \sum_{s_2=1}^{n} r_{s_1,s_2,z_1} \times \text{Prob}[X_t = s_1, X_{t-1} = s_2] \tag{4}$$

where $\text{Prob}[X_{t-1} = s_2, Y_t = z_1]$ is the joint probability distribution of X_{t-1} and Y_t, and $\text{Prob}[X_t = s_1, X_{t-1} = s_2]$ is the joint probability distribution of X_t and X_{t-1}. Suppose our desired equilibrium/stationary distributions of sentences and relations, i.e., the SentTopic-MultiRank values of sentences and relations, are given by

$$\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n]^T \text{ and } \bar{\mathbf{y}} = [\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_K]^T$$

respectively, with

$$\bar{x}_{s_1} = \lim_{t\to\infty} \text{Prod}[X_t = s_1] \text{ and } \bar{y}_{z_1} = \lim_{t\to\infty} \text{Prod}[Y_t = z_1]$$

for $1 \le s_1 \le n$ and $1 \le z_1 \le K$.

As (3) and (4) are coupled together and they involve two joint probability distributions, a product form of individual probability distributions is employed to replace the joint probability distributions in (3) and (4) by assumption:

$$\text{Prob}[X_{t-1} = s_2, Y_t = z_1] = \text{Prob}[X_{t-1} = s_2]\text{Prod}[Y_t = z_1]$$
$$\text{Prob}[X_t = s_1, X_{t-1} = s_2] = \text{Prob}[X_t = s_1]\text{Prod}[X_{t-1} = s_2]$$

Hence, using the above assumptions and considering t goes to infinity, (3) and (4) becomes

$$\bar{x}_{s_1} = \sum_{s_2=1}^{n} \sum_{z_1=1}^{K} o_{s_1,s_2,z_1} \bar{x}_{s_2} \bar{y}_{z_1}, \quad s_1 = 1, 2, \cdots n \tag{5}$$

$$\bar{y}_{z_1} = \sum_{s_1=1}^{n} \sum_{s_2=1}^{n} r_{s_1,s_2,z_1} \bar{x}_{s_1} \bar{x}_{s_2}, \quad z_1 = 1, 2, \cdots K \tag{6}$$

Under the tensor operation for (5) and (6), we solve the following tensor (multivariate polynomial) equations:

$$\mathscr{O}\bar{\mathbf{x}}\bar{\mathbf{y}} = \bar{\mathbf{x}} \text{ and } \mathscr{R}\bar{\mathbf{x}}^2 = \bar{\mathbf{y}} \tag{7}$$

with

$$\sum_{s_1=1}^{n} \bar{x}_{s_1} = 1 \text{ and } \sum_{z_1=1}^{K} \bar{y}_{z_1} = 1 \tag{8}$$

to obtain the SentTopic-MultiRank values of sentences and relations. An efficient iterative algorithm to solve the tensor equations in (7) is summarized as Algorithm 1. With respect to theoretical analysis of this algorithm, please refer to (Ng et al., 2011) for details.

Algorithm 1: The MultiRank Algorithm

Input: Two tensors $\mathscr{O}$ and $\mathscr{R}$, two initial probability distributions $\mathbf{x}_0$ and $\mathbf{y}_0$ ($\sum_{s_1=1}^{n}[\mathbf{x}_0]_{s_1}=1$ and $\sum_{z_1=1}^{K}[\mathbf{y}_0]_{z_1}=1$) and the tolerance ϵ
Output: Two stationary probability distributions $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$
Procedure:

1. Set $t=1$;
2. Compute $\mathbf{x}_t=\mathscr{O}\mathbf{x}_{t-1}\mathbf{y}_{t-1}$;
3. Compute $\mathbf{y}_t=\mathscr{R}\mathbf{x}_t^2$;
4. If $\| \mathbf{x}_t-\mathbf{x}_{t-1} \|+\| \mathbf{y}_t-\mathbf{y}_{t-1} \|<\epsilon$, then stop, otherwise set $t=t+1$ and goto Step 2.

In Algorithm 1, the setting of initial probability distributions does not influence the finally generated stationary probability distributions. We simply set uniform distribution for those sentences while with regard to topics, we set their initial distribution using the topic distribution of the entire corpus. Obviously, above algorithm has nothing to do with any query, so it does not apply to biased summarization task, which is further discussed in subsequent section.

3.4 Prior configuration in query-focused multi-document summarization

Query summarization requires that generated summary could not only express the salient content of target document collection, but also bias to the information needs of query. Then, how to exert the query information to the MultiRank process? Motivated by the idea of topic-sensitive PageRank (Haveliwala, 2003) and random walk with restart (Tong et al., 2008), we consider this issue by assigning desired limiting probability distributions towards sentences and topics, respectively. More specifically, we modify the tensor equations in (7) as follows:

$$(1-\mu)\mathscr{O}\bar{\mathbf{x}}\bar{\mathbf{y}}+\mu\mathbf{o}=\bar{\mathbf{x}} \tag{9}$$

$$(1-\nu)\mathscr{R}\bar{\mathbf{x}}^2+\nu\mathbf{r}=\bar{\mathbf{y}} \tag{10}$$

with Equation 8, where $\mathbf{o}$ and $\mathbf{r}$ are two devised prior probability distributions for sentences and topics/relations, respectively. Here, ($\sum_{i=1}^{n}[\mathbf{o}]_i=1$ and $\sum_{j=1}^{K}[\mathbf{r}]_j=1$), and $0\le\mu,\nu<1$, are two parameters for controlling the influence degree of two prior probability distributions. In experiments, we set $\mu=0.5$ and $\nu=0.9$ simply as (Xutao et al., 2012) indicated. Accordingly, the new iterative algorithm could be presented as Algorithm 2.

Noting that (Xutao et al., 2012) has ever provided a similar solution framework for their proposed HAR model, a modified version of MultiRank on the basis of HITS's principle, to cope with query search. Nonetheless, for one thing, our Algorithm 2 is designed specifically for MultiRank; for another, we further explicitly give the expressions of $\mathbf{o}$ and $\mathbf{r}$ for our summarization application in following sections.

3.4.1 Sentence prior

We try to normalize the relatedness of sentences towards the query description as those sentences' prior distribution. The most commonly used method of computing sentence relatedness is cosine measure. Nevertheless, we opt for an approach that is similar with what adopted by

Algorithm 2: The MultiRank Algorithm Integrated with Prior

Input: Two tensors $\mathscr{O}$ and $\mathscr{R}$, two initial probability distributions $\mathbf{x}_0$ and $\mathbf{y}_0$ ($\sum_{s_1=1}^{n}[\mathbf{x}_0]_{s_1}=1$ and $\sum_{z_1=1}^{K}[\mathbf{y}_0]_{z_1}=1$), two prior distributions of sentences and topics $\mathbf{o}$ and $\mathbf{r}$, two weighting parameters $0\leq\mu,\nu<1$ and the tolerance ϵ
Output: Two stationary probability distributions $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$
Procedure:

1. Set $t=1$;
2. Compute $\mathbf{x}_t$=(1-μ)$\mathscr{O}\mathbf{x}_{t-1}\mathbf{y}_{t-1}+\mu\mathbf{o}$;
3. Compute $\mathbf{y}_t$=(1-ν)$\mathscr{R}\mathbf{x}_t^2+\nu\mathbf{r}$;
4. If $\|\mathbf{x}_t-\mathbf{x}_{t-1}\|+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|<\epsilon$, then stop, otherwise set $t=t+1$ and goto Step 2.

(Kurland and Lee, 2010) as well as (Erkan, 2006). They all defined the sentence relatedness (e.g., u and v) as the generation probability of u given v or of v given u. Naturally, sentence-word distribution, as an initial step, must be acquired. They assumed that the weight of a word in certain sentence mainly depends on the word's occurrence frequency in that specific sentence. Given a sentence s, a straightforward way of computing the weights of words contained in s is to normalize their term frequency (TF):

$$P_{TF}(w|s)=c_s(w)/|s| \tag{11}$$

where $c_s(w)$ is times of word w occurring in s, and $|s|$ is the total number of words in s.

However, in our perspectives, it should not be neglected that different words have different degrees of information and instructions, no matter for a document or a sentence. For example, three sentences $s_1=(A,C)$, $s_2=(A:0.9,B:0.1)$ and $s_3=(D,B)$, where A, B, C and D denote 4 discriminative words, and 0.9, 0.1 mean the weights of A and B in the second sentence s_2. Now, if we are asked to distinguish which one of s_1 and s_3 has more similarity with s_2, apparently we should choose s_1. Even though the occurrence frequencies of A and B are the same in sentence s_2 (i.e., their generation probabilities are equal according to Equation 11), they actually have different importance towards s_2. As a result, s_1 is supposed to be more similar with s_2 because it has content s_2 more concerns.

With model LDA, we propose to derive the weight of word w with respect to sentence s where w occurs as follows:

$$P'(w|s)=\frac{P(w|C)}{\sum_{w'\in s}P(w'|C)} \tag{12}$$

where C means the entire corpus, and $P(w|C)$ could be easily acquired by combining corpus-topic distribution and topic-word distribution which are both basic outputs of LDA. To account for the unseen words for s, we smooth the above equation like this:

$$P(w|s)=(1-\lambda)P'(w|s)+\lambda P(w|C) \tag{13}$$

λ is a trade-off parameter and is set to 0.7 in experiments. Consequently, we can talk about

the generation probability of sentence u given another sentence v as:

$$\text{gen}(u|v) = P(u|v)^{\frac{1}{|u|}} = [\prod_{w \in u} P(w|v)]^{\frac{1}{|u|}} \tag{14}$$

We use the length of u (i.e., $|u|$) to conduct normalization so as to avoid the tendency that longer sentences get smaller generation probabilities. Moreover, we use a normalized generation probability to indicate the initial bias of sentences towards the query q, i.e.,

$$[\mathbf{o}]_{s_1} = \frac{\text{gen}(q|s_1)}{\sum_s \text{gen}(q|s)} \quad 1 \le s_1 \le n \tag{15}$$

We mark cosine measure as $cosine-based$, mark Equation 11 based generation probability method as $TF-based$, and label our proposed generation probability approach based on Equation 12 as $weight-based$. In the experimental part, they will be compared with each other to investigate their performance in constructing sentence prior.

3.4.2 Topic prior

When it turns to prior distribution of topics, we must first get the topic distribution of query description, i.e., the conditional probability $P(z_1|q)$, so

$$[\mathbf{r}]_{z_1} = P(z_1|q) = \frac{1}{|q|} \sum_{w \in q} P(z_1|w) \quad 1 \le z_1 \le K \tag{16}$$

where $|q|$ means the word number in q, and $P(z|w)$ is derived via Bayesian rule in LDA.

In experiments, we also set the two initial probability distributions in Algorithm 2 to equal with their corresponding initial distribution in Algorithm 1. Illustration of SentTopic-MultiRank with prior beliefs $\mathbf{o}$ and $\mathbf{r}$ is given as Figure 1(c).

4 Summary generation

After ranking process, sentences usually have close values if they own similar content. Consequently, we perform the modified MMR algorithm, proposed in (Wan et al., 2007), during sentence selection to control the information redundancy of generated summary. The principle of MMR algorithm is to reduce the ranking scores of remaining sentences if they are detected to have a degree of similarities with those sentences having been selected to construct a summary. The algorithm goes as follows:

1. Initiate two sets A=∅, B={s_i|i=1,2, ⋯, n}, and each sentence's initial value is set to its SentTopic-MultiRank score obtained in above section, i.e., $value(s_i)$=$\bar{x}_i$.
2. Sort the sentences in B by their current values in descending order.
3. Suppose s_i is the highest ranked sentence, i.e., the first sentence in B. Move sentence s_i from B to A and update the values of the remaining sentence(s) in B as follows:
   For each sentence s_j in B:
   $$value(s_j) = value(s_j) - \omega \cdot \tilde{S}_{ij} \cdot value(s_i) \tag{17}$$
   where ω is set to 5 in our experiments, $\tilde{S}$ is the normalized sentence similarity matrix.
4. Go to step 2 and iterate until $|B|$=0.

After the final scores are obtained for all sentences, several sentences with highest ranking scores are chosen to produce a summary until length limit is reached.

5 Experiments

5.1 Data sets and evaluation metrics

We validate our algorithm framework in both generic and query-biased multi-document summarization. For generic task, we conduct experiments on the data sets DUC2002[2] and DUC2004[3] in which generic multi-document summarization has been one of the fundamental tasks (i.e., task 2 in DUC2002 and task 2 in DUC2004). For query-related task, experiments are based on the main tasks of DUC2005[4] and DUC2006[5]. Each task has a gold standard data set consisting of document sets and reference summaries. Table 1 gives a short summary of above data sets. Documents are pre-processed by segmenting sentences and splitting words. Stop words are removed and the remaining words are stemmed using Porter stemmer[6].

	DUC2002	DUC2004	DUC2005	DUC2006
Task	Task2	Task2	the only task	the only task
Number of documents	567	500	1593	1250
Number of clusters	59	50	50	50
Data source	TREC	TDT	TREC	TREC
Summary length	200 words	665 bytes	250 words	250 words

Table 1: Summary of data sets

We use the ROUGE (Lin, 2004) (version 1.5.5) toolkit[7] for evaluation, which is officially adopted by DUC for evaluating automatic generated summaries. Here we report the average F-measure scores of ROUGE-1, ROUGE-2 and ROUGE-SU4, which base on Uni-gram match, Bi-gram match, and unigram plus skip-bigram match with maximum skip distance of 4 between the candidate summary and the reference summary, respectively.

5.2 System comparison

5.2.1 Generic multi-document summarization

As for generic multi-document summarization, we compare our proposed method SentTopic-MultiRank with following algorithms. (1)Lead Baseline: The lead baseline takes the first sentences one by one in the last document in a document set, where documents are assumed to be ordered chronologically. (2)Random: The method selects sentences randomly for each document collection. (3)DUC Best: The system with highest ROUGE scores among all the systems submitted. (4)LexPageRank: The method first constructs a sentence connectivity graph based on traditional cosine similarity and then conducts PageRank algorithm to determine global importance scores of sentences (Erkan and Radev, 2004). (5)ToPageRank: The method, proposed in (Pei et al., 2012), decomposes traditional PageRank into multiple PageRank via topic decomposition for generic multi-document summarization.

Tables 2-3 show the experimental results. From those statistics, we observe that ToPageRank is better than LexPageRank while our SentTopic-MultiRank outperforms all the competitors.

[2]http://www-nlpir.nist.gov/projects/duc/data/2002_data.html
[3]http://www-nlpir.nist.gov/projects/duc/data/2004_data.html
[4]http://www-nlpir.nist.gov/projects/duc/duc2005/tasks.html
[5]http://www-nlpir.nist.gov/projects/duc/duc2006/tasks.html
[6]http://tartarus.org/ martin/PorterStemmer/
[7]http://www.isi.edu/licensed-sw/see/rouge/

Systems	ROUGE-1	ROUGE-2	ROUGE-SU4
Lead	0.39860	0.16042	0.20315
Random	0.38469	0.11705	0.18007
LexPageRank	0.47473	0.22484	0.25850
DUC Best	0.49869	0.25229	0.28406
ToPageRank	0.49923	0.25735	0.29630
SentTopic-MultiRank	**0.50491**	**0.26714**	**0.30731**

Table 2: *F*-measure comparison on DUC2002

Systems	ROUGE-1	ROUGE-2	ROUGE-SU4
Lead	0.33182	0.06348	0.10582
Random	0.31857	0.06269	0.11780
LexPageRank	0.37875	0.08354	0.12770
DUC Best	0.38279	0.09216	0.13349
ToPageRank	0.40251	0.09555	0.14027
SentTopic-MultiRank	**0.41016**	**0.09917**	**0.14324**

Table 3: *F*-measure comparison on DUC2004

It suggests that: (1) the decomposition and analysis of topic information indeed benefit the improvement of summary quality, as ToPageRank and SentTopic-MultiRank systems perform. It might result from the fact that not only does target corpus usually involve multiple topics, but a user commonly wishes to obtain information from various aspects by reading a concise text. Analyzing corpus on topic level could deeply and comprehensively identify user desired content. (2)ToPageRank focuses on sentence interaction within a topic's range while SentTopic-MultiRank takes three kinds of relations (sentence-sentence, sentence-topic, topic-topic) into consideration wholly. Hence, propagation of impacts among sentences and topics (noting that topics are treated as relation types) is a bonus to determine the saliency score of sentences.

5.2.2 Query-biased multi-document summarization

To validate SentTopic-MultiRank in query-biased task, following systems are implemented as baselines. (1)Random: The same baseline as that in above generic task. (2)Manifold: ranking the sentences according to the manifold ranking scores. (3)Biased LexRank: A modified version of traditional random walk with prior belief. It was presented in (Erkan, 2006) and the sentence prior knowledge was denoted by sentence's similarity to the query description. (4)top three systems with the highest ROUGE scores that participated in the DUC2005 (S4, S15, S17) and the DUC2006 (S12, S23, S24) for comparison, respectively. Tables 4 and 5 present the performance of these systems on DUC2005 and DUC2006 data sets, respectively.

Systems	ROUGE-1	ROUGE-2	ROUGE-SU4
Random	0.30821	0.03976	0.10625
Biased LexRank	0.37324	0.07113	0.12805
Manifold	0.37497	0.07423	0.12907
S17	0.36933	0.07286	0.12937
S4	0.37584	0.07063	0.12868
S15	0.37656	0.07244	0.13248
SentTopic-MultiRank	**0.38803**	**0.07994**	**0.13532**

Table 4: *F*-measure comparison on DUC2005

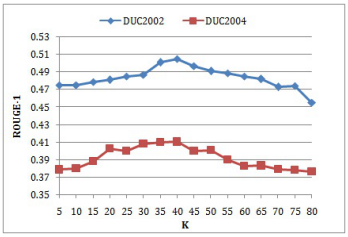
Systems	ROUGE-1	ROUGE-2	ROUGE-SU4
Random	0.34821	0.05297	0.11908
Biased LexRank	0.36814	0.08074	0.12993
Manifold	0.38867	0.08308	0.13307
S23	0.40973	0.09785	0.16162
S12	0.41053	0.09633	0.16074
S24	0.41081	0.09957	0.15248
SentTopic-MultiRank	**0.41567**	**0.10125**	**0.16431**

Table 5: F-measure comparison on DUC2006

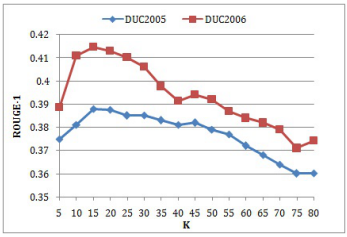
Apparently, our approach SentTopic-MultiRank has the optimal performance, especially outperform the two representative graph-based methods: Biased LexRank and Manifold. Noting that both the sentence networks of Biased LexRank and Manifold are only based on sentence similarity under a single relation type assumption, so the estimation of their sentence ranking scores is limited compared with SentTopic-MultiRank which skillfully considers query-biased information as well as the interactions between sentences and topics.

5.3 Influence of topic number K

LDA is a crucial tool in devising SentTopic-MultiRank and topic number K also denotes the amount of different relation types. In order to investigate how the topic number K exerts influence to system performance, K is varied from 5 to 80. Figures 2(a)-2(b) show the influence of K over ROUGE-1 with respect to generic and query-oriented multi-document summarization, respectively. On the whole, the evaluation scores rise with the increase of K, and reach their respective maximum, then gradually go down when K continues to become larger. Despite some slight fluctuations, they don't matter to the overall conclusion. We could draw the following two meaningful conclusions. (1)It indicates that topic number indeed influences the algorithm result. In fact, our system would be similar with LexPageRank if $K = 1$. (2)Interestingly, two tasks do not reach their optimal situations at an approximate K value, more exactly, generic task requires a relatively large K ($K \approx 40$) while query-biased prefers small K ($K \approx 15$) (here our adoption of an approximation for K is because that we actually vary K every 5 from the start value $K = 5$ in experiments). It may derive from the fact that a query description usually involves few topics, whereas, when applying SentTopic-MultiRank to generic task, we could image the whole corpus as a big query, which concerns about more topics.



(a) ROUGE-1 vs. K for generic multi-document summarization



(b) ROUGE-1 vs. K for query-focused multi-document summarization

Figure 2: Influence of K

5.4 Comparison of three sentence prior methods

In Section 3.4.1, we named *weight-based* for our proposed sentence prior method. In order to confirm its advantage in acquiring more favorable sentence prior distribution, we compare it with two competitors: *cosine-based* and *TF-based* using query-biased summary qualities on DUC2005 and DUC2006. Section 5.3 has indicated that when K is about 15 the quality of query-oriented summary is optimal. For simplicity, we compare the three methods using ROUGE-1 metric under the same condition $K = 15$. Figure 3 gives comparison results. For the space limit, here we give up showing the comparison results on ROUGE-2 and ROUGE-SU4.

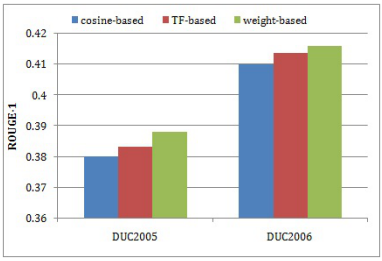


Figure 3: Comparison of three sentence prior methods, $K = 15$.

As Figure 3 shows, generation probability methods, including *TF-based* and *weight-based*, both outperform *cosine-based*. In addition, system using *weight-based* method to generate sentence prior indeed produces summaries with the highest quality. It not only suggests the effectiveness of generation probability methods, but also points out that our method, considering term weights instead of term frequencies to acquire generation probabilities of words given a sentence, is more useful to derive sentence affinity.

Conclusion and perspectives

In this study, we exploit LDA to devise a novel model named SentTopic-MultiRank based on MultiRank algorithm for multi-document summarization. Our intention is to map sentence relatedness over various latent topics to heterogeneous network where multiple topics are treated as various kinds of relations. We apply our approach to both generic and query-biased multi-document summarization. In generic task, we use MultiRank algorithm to directly determine the importance of all sentences, while in query-sensitive task, some query-biased prior beliefs are added to sentences and topics of the SentTopic-MultiRank network, so that some top-ranked sentences are supposed to not only demonstrate the core content of target corpus, but also bias to the information needs of a user's query. Extensive experiments have been performed to prove the effectiveness of our approach in improving the summary quality.

In future work, for one thing, we may apply our model to new summarization task, such as updated summarization; for another, integrating manifold-ranking process with SentTopic-MultiRank might be an interesting idea in some query-related applications.

Acknowledgments

This work is financially supported by Grant SKLSDE-2010KF-03 and NSFC Grant 60933004.

References

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Erkan, G. (2006). Using biased random walks for focused summarization. In *Proceedings of the 2006 Document Understanding Conference*.

Erkan, G. and Radev, D. (2004). Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*, volume 4.

Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128. ACM.

Haveliwala, T. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Kolda, T. and Bader, B. (2006). The tophits model for higher-order web link analysis. In *Workshop on Link Analysis, Counterterrorism and Security*, volume 7, pages 26–29.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM.

Kurland, O. and Lee, L. (2010). Pagerank without hyperlinks: Structural reranking using links induced by language models. *ACM Transactions on Information Systems (TOIS)*, 28(4):18.

Lin, C. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (WAS 2004)*, volume 16.

Lin, C. and Hovy, E. (2002). From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 457–464. Association for Computational Linguistics.

Liu, Y., Zhong, S., and Li, W. (2012). Query-oriented multi-document summarization via unsupervised deep learning. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Liu, Z., Huang, W., Zheng, Y., and Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366–376. Association for Computational Linguistics.

Manning, C., Schutze, H., and MITCogNet (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

Mihalcea, R. and Tarau, P. (2005). A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP*, volume 5.

Ng, M., Li, X., and Ye, Y. (2011). Multirank: co-ranking for objects and relations in multi-relational data. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1217–1225. ACM.

Nie, L., Davison, B., and Qi, X. (2006). Topical link analysis for web search. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 91–98. ACM.

Ouyang, Y., Li, S., and Li, W. (2007). Developing learning strategies for topic-based summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 79–86. ACM.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

Pei, Y., Yin, W., Zhang, F., and Huang, L. (2012). Generic multi-document summarization using topic-oriented information. In *Proceedings of the 2012 Pacific Rim International Conference on Artificial Intelligence*, pages 435–446.

Radev, D., Jing, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Saggion, H., Bontcheva, K., and Cunningham, H. (2003). Robust generic and query-based summarisation. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 235–238. Association for Computational Linguistics.

Shen, C. and Li, T. (2011). Learning to rank for query-focused multi-document summarization. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 626–634. IEEE.

Shen, D., Sun, J., Li, H., Yang, Q., and Chen, Z. (2007). Document summarization using conditional random fields. In *Proceedings of IJCAI*, volume 7, pages 2862–2867.

Tong, H., Faloutsos, C., and Pan, J. (2008). Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346.

Wan, X., Yang, J., and Xiao, J. (2007). Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI*, volume 7, pages 2903–2908.

Xutao, L., Michael, K., and Yuming, Y. (2012). Har: Hub, authority and relevance scores in multi-relational data for query. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 141–152.

Zhou, D., Weston, J., Gretton, A., Bousquet, O., and Schlkopf, B. (2004). Ranking on data manifolds. *Advances in neural information processing systems*, 16:169–176.

Zhou, L. and Hovy, E. (2003). A web-trained extraction summarization system. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 205–211. Association for Computational Linguistics.