Knowledge-Based Systems 46 (2013) 33–42

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Exploiting relevance, coverage, and novelty for query-focused multi-document summarization

Wenjuan Luo [a,b,*], Fuzhen Zhuang [a], Qing He [a], Zhongzhi Shi [a]

[a] The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 7 June 2012
Received in revised form 26 February 2013
Accepted 28 February 2013
Available online 20 March 2013

Keywords:
Query-focused document summarization
Dependent relevance
Coverage
Novelty
PHITS

ABSTRACT

Summarization plays an increasingly important role with the exponential document growth on the Web. Specifically, for query-focused summarization, there exist three challenges: (1) how to retrieve query relevant sentences; (2) how to concisely cover the main aspects (i.e., topics) in the document; and (3) how to balance these two requests. Specially for the issue relevance, many traditional summarization techniques assume that there is independent relevance between sentences, which may not hold in reality. In this paper, we go beyond this assumption and propose a novel Probabilistic-modeling Relevance, Coverage, and Novelty (PRCN) framework, which exploits a reference topic model incorporating user query for **dependent relevance** measurement. Along this line, topic coverage is also modeled under our framework. To further address the issues above, various sentence features regarding **relevance** and **novelty** are constructed as features, while moderate topic **coverage** are maintained through a greedy algorithm for topic balance. Finally, experiments on DUC2005 and DUC2006 datasets validate the effectiveness of the proposed method.

1. Introduction

Automatic document summarization is the process of generating a textual summary from the all-embracing corpus of documents [24]. Motivated by the prevalence of WWW, there is a critical need for summarization techniques that can effectively process large-scale text documents accumulated on the Web [10]. A promising direction to overcome this obstacle is query-focused summarization, which aims to create summaries that are relevant to a given query or user profile [1,21].

Pointed out in [39], a good query-focused summary should be relevant to the given query and preserve the information contained in the documents as much as possible. Furthermore, the summary should be able to maintain moderate balance among different aspects (which is the same as topics in this paper) with least redundancy [37,38]. To that end, there are three challenges: (1) how to retrieve query relevant sentences; (2) how to concisely cover the main aspects in the document; and (3) how to balance these two requests.

(1) As for query *relevance*, which is central to many text retrieval problems [30,36,11,19,4], this measurement determines whether or not a summary accounts for the interests of user profile (or given query). Many existing summarization techniques assume that the relevance of a sentence is independent of the relevance of other sentences [40]. Unfortunately, in reality, the utility of selecting one sentence, in general, may depend on which sentences the user has already seen [44,5]. In this study, we go beyond the independent relevance assumption and model sentence relevance in a probabilistic manner.
(2) As for *coverage* [20,21] of document contents, which is another target of summarization, this purpose demands summaries to cover the main aspects in the document as much as possible [39]. In addition, the final summary should maintain appropriate aspect coverage distribution in consistent with the original documents. Thereafter, we impose topic *balance* [21] in the representative sentence extraction process for rational coverage of document topics.
(3) Note that for a summary of limited length, *relevance* and *coverage* are two requirements that may contradict each other. In other words, If the summarization process overemphasizes either requirement, it may not be able to meet the other requirement. To resolve the conflict, one may resort to the *novelty* [45] for novel information richness conditioned that the summary is already relevant to the given topic.

* Corresponding author at: The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.

E-mail addresses: luowj@ics.ict.ac.cn (W. Luo), zhuangfz@ics.ict.ac.cn (F. Zhuang), heq@ics.ict.ac.cn (Q. He), shizz@ics.ict.ac.cn (Z. Shi).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.02.015

In this paper, we propose a Probabilistic-modeling Relevance, Coverage, and Novelty framework (denoted as PRCN) based on the PLSA (Probabilistic Latent Semantic Analysis) [14] and the PHITS (Probabilistic Hyperlink-Induced Topic Search) [6], which considers the three issues above (i.e., relevance, coverage and novelty) for query-focused multi-document summarization. Without the independent relevance assumption, we model relevance and coverage in a probabilistic manner, while novelty and other sentence features are constructed according to equations we defined later in the paper. Ultimately, a greedy sentence selection method is developed for topic balance, where suitable topic proportions are maintained.

In summary, our work contributes in the following perspectives: (1) A novel joint probabilistic framework PRCN is proposed for relevance, coverage, and novelty acquisition; (2) sentence features regarding **document relevance/query relevance** and **document novelty/query novelty** are constructed and quantitatively estimated; and (3) **topic balance/query balance** are imposed in a greedy balance algorithm based on the constructed sentence features for summary sentence selection.

Finally, experiments have been performed on the DUC2005 and DUC2006 benchmark datasets to investigate the effectiveness of our approach. And the results demonstrate that PRCN can outperform other methods on the ROUGE-2 and ROUGE-W [23] evaluation metrics with a significant margin.

1.1. Overview

The remainder of the paper is organized as follows: we first provide a detailed amount of related work in Section 2, and then we formally define the problem of query-focused multi-document summarization and propose our solution in Section 3. The EM derivation of our model is detailed in Section 4. The sentence feature definitions and quantifications are given in Section 5, where a greedy algorithm for sentence selection is also given. Experimental evaluation is shown in Section 6. Finally, Section 7 concludes our paper.

2. Related work

2.1. Query-focused multi-document summarization

According to different categorization criteria, summarization techniques can be categorized into abstract-based and extract-based (reproducing sentence or not), multi-document and single-document (more than one document or not), query-focused and generic (given query or not), supervised and unsupervised (with training set or not) methods [20]. In this paper, we focus on unsupervised, extract-based, query-focused, multi-document summarization.

Query-focused multi-document summarization aims to distill the most important information from a set of documents to generate a compressed summary that is relevant to a given query. Shen and Li [32] propose a principled and versatile framework using the minimum dominating set, which could be categorized into the graph-based methods. Along this line, Wan et al. utilize the manifold-ranking algorithm for sentence ranking [39,37], where a graph is generated for document representation. However, as for graph-based methods, the performance may change as the similarity measurement varies.

Meanwhile, supervised extractive approaches are also employed for summarization, such as Support Vector Machine (SVM) [13], Logistic Regression (LR) model [35], Bagging and Gaussian Process (GP) [15]. As pointed out, such classifiers or regression models assume independent relevance between sentences without leveraging the internal relationship among sentences [15]. On the other hand, Hidden Markov Model (HMM) [8] based methods attempt to break this assumption, however, in reality, HMMs suffer from the problem of overfitting in the test data. Specially, when the feature space is large and the features are independent, the training process of HMM becomes intractable [33].

From another angle, feature-construction based sentence ranking methods are also developed for summarization, such as structural SVM with three types of constraints proposed in [20], where diversity, coverage, and balance are enhanced for summary generation. Considering the given topic, Li et al. [21] propose a sentence ranking probability model that incorporates novelty, coverage, and balance with respect to a given topic. Besides, traditional sentence features such as sentence length, sentence position and similarity to title are listed as [27,33].

Along this line of work, we also construct a set of features for sentence salience ranking. Yet our study makes unique contributions by dependently modeling relevance and coverage, which generates probabilities for subsequent feature qualification. Incidently, we define and distill sentence features in a different manner from the above mentioned feature-based methods. Besides, we propose a greedy sentence selection algorithm for rational coverage on different topics.

2.2. Relevance ranking

Sentence ranking is one of the issues of most concern for extractive summarization [41]. For query-focused summarization, relevance measures how relevant a sentence is to the given query, which has been extensively investigated as an important factor [1,18,12] in information retrieval. Traditional retrieval techniques [46,34] assume that the relevance of a sentence is independent of the relevance of other sentences. The notion of independent relevance ranking is firstly mentioned in [44], which, in our case, means that the utility of a sentence in ranking is dependent on other sentences in the ranking list.

A recent study [12] discover that on average 57% of the sentences in the document are query-relevant and that an ideal query expansion leads to a situation in which almost all the sentences in the input become relevant. Therefore, such a vast sea of potentially relevant sentences may be highly redundant with each other or (extremely) contain duplicate information [3]. In the process of summarization, later sentences should supplement earlier selected sentences, rather than redundantly deliver similar content again and again [5].

For this dependent relevance analysis, Zhai et al. [44] propose a mixture model combined with query likelihood relevance ranking, while Clarke et al.[5] develop a new framework and obtained a specific evaluation measure based on cumulative gain. On the whole, the manner of dependent relevance modeling aims to understand documents with consideration of internal relationship between sentences and consequently eliminate redundancy implicitly.

In this study, we dependently model sentence relevance in regard to topics under a joint probabilistic framework, where sentence relevance are analyzed at the aspect level. So far as we know, our method is the first to decompose sentence relevance into the topic space for more internal relationship understanding in query-focused multi-document summarization. Note that we do not presume that sentence relevance are independent with each other, but dependently model it in a probabilistic manner.

2.3. Coverage modeling

Coverage focuses on avoiding the loss of main information in the documents, which is an important issue in summarization

[20]. As it is pointed out, most state-of-art approaches [17,33,26] employ 0–1 loss functions to quantify the coverage of sentences. However, such measurement roughly computes coverage and ignores significant information at the aspect level. From another angle, other methods obtain coverage through sentence clustering and assign to each sentence with a selection preference according to the cluster distribution [29,43,21].

In this study, we enforce coverage through modeling sentences with respect to a set of topics, which produces sentence coverage (or say, probability) over each aspect. Note that we do not simply assign each sentence to any single topic but employ a probability distribution for more reasonable coverage allocation over aspects. Our work differs from the previously mentioned work in that we model sentence coverage under a joint probabilistic framework.

Moreover, we argue that the final summary should not only cover the main content as much as possible, but also wisely cover content aspects in proportion to the original aspect distributions in the documents. So far as we know, our work makes novel contribution to assigning reasonable (balanced) coverage to document content based on probabilities at the aspect level.

2.4. Novelty acquisition

There are some empirical study for novelty criteria in [5,20,29,21,44,42]. The widely used MMR (Maximal Marginal Relevance) ranking method proposed by Carbonell and Goldstein [3] selects sentences according to a combined criterion of query relevance and novelty of information. The novelty is defined to measure the degree of dissimilarity between the sentence being considered and previously selected ones in the summary. The need for "relevant novelty" was motivated as a potentially superior criterion for sentence ranking, which could be formulated as [3]:

$$\mathrm{MMR} = \arg\max_{D_i \in R \backslash S}[\lambda(Sim_1(D_i, \mathcal{Q})) - (1-\lambda)\max_{D_j \in S} Sim_2(D_i, D_j)], \tag{1}$$

where $\mathcal{Q}$ is a query or user profile, $R \backslash S$ is the subset of documents yet unselected in R, D_i is the ith document in the subset, Sim_1 and Sim_2 are different similarity measurements. Note that the MMR model acquires information novelty through eliminating redundancies between candidate sentences and selected sentences. Furthermore, Wan et al. [39] explicitly define the "information novelty" to measure the novelty degree of one sentence with respect to other sentences, i.e., the dissimilarities between the sentence being considered and all other sentences.

With consideration of "query (or user profile) novelty", Li et al. [21] define the "Topic-Aware Novelty", which penalizes word redundancies in view of the relevance between words and query. Zhai et al. [44] develop a reference topic model for novelty measurement. Specially, Clarke et al. clear that novelty is the need to avoid redundancy and developed a probabilistic framework for novelty acquisition [5].

In this study, we define the novelty of a sentence as some evaluation metric that would indicate how much novel information the sentence contains. As you may notice, all the previously mentioned work obtain novelty either from the document perspective or from the query perspective, in this paper, we construct novelty regarding both query novelty and document novelty.

2.5. PLSA, PHITS, and LDA

PRCN (Probabilistic-modeling Relevance, Coverage, and Novelty) is a joint model of the PLSA (Probabilistic Latent Semantic Analysis) [14] and the PHITS (Probabilistic Hyperlink-Induced Topic Search) [6], which are two typical topic models. Typical topic models such as PLSA and LDA (Latent Dirichlet Allocation) [2] are attracting growing interests for text mining. Specifically, PLSA is a probabilistic model for modeling the document and word co-occurrence for topic generation, which provides a probabilistic understanding of words and documents from the latent topic space. PHITS is proposed as a joint probabilistic model for modeling the inter-connectivity of document collections, which performs a probabilistic factoring of document citations used for bibliometric analysis. While LDA presumes some Dirichlet priors for documents and words, and thus performs more complex inference in parameter and topic distribution estimation. LDA has a better statistical foundation by defining the topic-document distribution, it allows inferencing on new documents and better suffers the problem of overfitting [22], where both are known as the deficits of PLSA based methods. However, in this paper, we do not perform relevance analysis based on LDA, and only implement the classical LDA in [2] as a baseline for comparison in the experiments.

Particularly, PHITS in [7] models word occurrences and citations/links into "topic" factors, with α as the relative importance assigned to predicting terms:

$$\mathcal{L} = \sum_j \left\{ \alpha \sum_i \frac{N_{ij}}{\sum_{i'} N_{i'j}} \log \sum_k P(t_i|z_k)P(z_k|d_j) \right. \\ \left. + (1-\alpha) \sum_l \frac{A_{lj}}{\sum_{l'} A_{l'j}} \log \sum_k P(c_l|z_k)P(z_k|d_j) \right\}, \tag{2}$$

where N_{ij} is the word occurrences of word t_i in document d_j, $a_{l,j}$ is the citation link between content c_l and document d_j. The normalization by term/citation counts ensures that each document is given the same weight in the decomposition, regardless of the number of observations associated with it.

In this paper, PRCN considers each sentence as a unit and makes unique contribution in adopting the sentence similarity as the intra-sentence link (citation in PHITS) for model generation. It is worth mentioning that PHITS and PRCN do not suffer from the problem of stability [28] as HITS (Hyperlink-Induced Topic Search) [16] does. Since PHITS projects the citations (similarities for PRCN) into a latent topic space that best describes the citations, the optimization process of PHITS aims to discover such subspace, where similar contents are gathered by the same topic. With a different focus, the HITS algorithm strives to find the authority nodes through iterations of similarity propagation and weight dissemination. As is pointed out [28], if the output of an algorithm is a subspace, then the stability considerations may not be a matter of primary concern, such as LSI [9]. As a matter of fact, PLSA is the probabilistic LSI, and thence would not suffer the problem of stability. As PHITS and PRCN are both PLSA-based methods, thereafter neither of them would encounter the problem of stability.

3. Problem definition and solution

In this section, we first list the frequently used notations in Table 1 and then introduce the problem together with our solution. Given a user query $\mathcal{Q}$ and a set of sentences $\mathcal{S} = \{s_1, s_2, \ldots, s_n\} \subset R^m$, where s_i represents the ith sentence in the document, R^m denotes the whole corpus. Our goal is to select $\mathcal{Q}$-relevant sentences with highest R_i as summary:

$$\mathfrak{S} = \arg\max_{s_i}\{R_i | s_i \text{ is } \mathcal{Q} \text{ relevant}\}. \tag{3}$$

Note that in the sentence selection stage, we select each sentence with the highest R_i among the rest unselected sentences in the light of Eq. (3). Incidently, after one sentence is chosen, we apply a greedy algorithm to update the R_i of rest sentences, where the details would be presented in Section 5.

Specifically, our method is divided into three steps: (1) the joint probabilistic modeling framework is firstly utilized to dependently model sentence relevance and coverage on topics. (2) Afterwards, sentence features emphasizing relevance, coverage, and novelty

Table 1
Term notations.

Symbol	Description
$\mathbb{T}$	The term frequency matrix
$\mathbb{A}$	The similarity matrix
x_i	The ith sentence in summary
$\mathfrak{S}$	The final summary
s_i	The ith sentence in document
$\mathcal{S}$	The set of sentences
t_{ij}	The frequency of w_j in s_i
a_{ij}	The similarity between s_i, s_j
θ	All the no-query (content or background language) topics
$\mathcal{Q}$	The user query
θ_q	The topic describing the query $\mathcal{Q}$
R_i	The ranking score of s_i
Θ	The latent topic (aspect) space
Λ	All the parameters
θ'	All the topics $\{\theta, \theta_q\}$
α	The weight of modeling coverage

are constructed based on the generated probabilities. (3) Finally, a greedy sentence selection method is applied for summary generation.

3.1. Topic relevance modeling

For relevance modeling, we incorporate the cosine similarities between sentences as the preliminary sentence relevances for probabilistic modeling, which aims to analyze sentence relevances dependently at the topic level. Specifically, given sentence s_i and s_j, PRCN computes the similarity a_{ij} between s_i and s_j as:

$$a_{ij} = \begin{cases} \frac{s_i \cdot s_j}{\|s_i\| \cdot \|s_j\|}, & i \neq j, \\ 0, & \text{Otherwise}. \end{cases} \tag{4}$$

Given the similarity matrix $\mathbb{A}$ between sentences, with each element a_{ij} representing the cosine similarity between content s_i and sentence s_j, let Θ denote the latent topic space, to enforce dependent relevance modeling, we aim to maximize the following log likelihood function:

$$l(\mathbb{A}|\Lambda) = \sum_{\Theta} \log P(\mathbb{A}, \Theta|\Lambda) = \sum_{i,j} \frac{a_{ij}}{\sum_{i'} a_{i'j}} \log \sum_{\theta'} P(s_j|\theta')P(\theta'|s_i), \tag{5}$$

where Λ includes the parameters of $P(s_j|\theta')$ and $P(\theta'|s_i)$. Through the optimization of the above function, we aim to cluster sentences with similar contents together, where sentence relevance are modeled analytically in the topic space.

Note that through the above optimization process, sentences are represented as a vector of probabilities over a set of aspects, which makes it feasible to analyze dependent relevance in the topic space. The rationale behind this is, since the topics in the PHITS model are orthogonal, it becomes reasonable to analyze sentence relevance dependently inside topics and independently across topics. Without independent relevance assumption, sentence relevances are decomposed into vectors among the topic simplex, thus makes it feasible to analyze and quantify the dependent relevance.

3.2. Topic coverage modeling

To ensure that the latent topic space captures the words and sentences contained in the documents, we utilize the term-frequency matrix for topic coverage modeling. The joint probabilistic framework is adopted to model words and sentences, where through the clustering of words and sentences, distinct topics emerge with coverage (or say probability) on words and sentences. Given the term-frequency matrix $\mathbb{T}$ of sentences, with each element t_{ij} corresponding to the frequency of word w_j in sentence s_i, the log likelihood of the document is:

$$\begin{aligned} l(\mathbb{T}|\Lambda) &= \sum_{\Theta} \log P(\mathbb{T}, \Theta|\Lambda) \\ &= \sum_{i,j} \frac{t_{ij}}{\sum_{i'} t_{i'j}} \left\{ \log \sum_{\theta'} P(w_j|\theta')P(\theta'|s_i) \right\}, \end{aligned} \tag{6}$$

where Λ includes the parameters to be estimated. Topic coverage modeling clusters each word and each sentence into topics and thus the latent topics θ' cover all the information among words and sentences.

$P(w_j|\theta')$ and $P(\theta'|s_i)$ could be interpreted as the coverage of w_j on each topic and the coverage of each topic on s_i, respectively. Note that $\sum_j P(w_j|\theta') = 1$ indicates that the latent topics are covered by word clusters, while $\sum_{\theta'} P(\theta'|s_i) = 1$ implies that sentences are distributed among the latent topic space. The goal of topic coverage modeling is to produce topics that best represent word clusters while cover all the information among sentences.

The optimization of the above likelihood function will converge to the latent topic space where words or sentences belonging to the same aspect are covered by the same topic. The coverage modeling framework makes it possible to cluster words or sentences in a probabilistic manner, which also provides a probabilistic representation of coverage on words and sentences.

3.3. Reference topic model

The so-called reference topic model was firstly proposed in [44] for novelty and redundancy measurements beyond independent relevance. Suppose a generative mixture model for a new sentence, in which one component is the old reference topic model (in our case, the topic about the query θ_q) and the other components are background language models (the content topics θ).

Given the observed new sentence $s_i = \{w_1, \ldots, w_n\}$, we estimate the mixing weights for the reference topic model (or the background models) as:

$$l(s_i, \lambda|\theta') = \sum_{j=1}^{n} \log \left\{ \lambda_q P(w_j|\theta_q) + \sum_{\theta} \lambda_\theta P(w_j|\theta) \right\}, \tag{7}$$

where $\lambda = \{\lambda_q, \lambda_\theta\}$ is a relevance vector with $\lambda_q + \sum_\theta \lambda_\theta = 1$, λ_q is the query relevance and the rest λ_θ are the document content relevances, respectively. The estimated weights can be interpreted as the extent to which the new sentence can be explained by the query topic as opposed to the background contents. From another point of view, λ_q can also serve as a measure of query redundancy as pointed out by Zhai et al. [44]. Also as is testified in their study, the reference topic model dependently analyzes sentence relevance, and more details could be referred to [44].

In our study, we incorporate the user profile as the reference topic so as to dependently model query relevance. Take a careful look at the reference topic model, compared to the conditional probability in the topic relevance modeling stage ($\log \sum_{\theta'} p(s_j|\theta')p(\theta'|s_i)$ in Eq. (5)) and coverage modeling stage ($\log \sum_{\theta'} p(w_j|\theta')p(\theta'|s_i)$ in Eq. (6)), the parameters λ_q and λ_θ could be respectively interpreted as $P(\theta_q|s_i)$ and $P(\theta|s_i)$ in the likelihood functions.

Thereafter, the user profile (query) could be dependently included in the relevance and coverage modeling stage, where user profile and document contents could be described by a set of aspects. The reference topic model orthogonally projects query and background content into the latent topic space, thus it becomes feasible to dependently analyze sentence query relevance. As the reference topic model could be easily generalized into our joint probabilistic model, query relevance and coverage would be dependently modeled.

3.4. Model combination

We combine the relevance modeling (Eq. (5)) and topic coverage modeling (Eq. (6)) with weight α, and specialize our model into the reference topic (Eq. (7)) model with user query:

$$\begin{aligned}\mathcal{L} &= \alpha \cdot l(\mathbb{T}|\Lambda) + (1-\alpha)\cdot l(\mathbb{A}|\Lambda)\\ &= \alpha \cdot \sum_{i,j} \frac{t_{ij}}{\sum_{i'} t_{i'j}} \log\left\{P(w_j|\theta_q)P(\theta_q|s_i) + \sum_{\theta} P(w_j|\theta)P(\theta|s_i)\right\}\\ &\quad + (1-\alpha)\cdot \sum_{i,j} \frac{a_{ij}}{\sum_{i'} a_{i'j}} \log\left\{P(s_j|\theta_q)P(\theta_q|s_i) + \sum_{\theta} P(s_j|\theta)P(\theta|s_i)\right\}, \qquad (8)\end{aligned}$$

where α denotes the relative weight of modeling the coverage. According to the reference topic model, we consider $P(\theta_q|s_i)$ as query relevance and $P(\theta|s_i)$ as document content relevance vector, respectively.

Note that under the specialized framework, the topic relevance and coverage are modeled regarding given query θ_q and background (i.e., content) topics. In the next section, we derive an EM algorithm to optimize the likelihood function. As a result, relevance and coverage are modeled and quantified dependently in our method. To be clear, novelty is implicitly considered in the reference model generation stage (the orthogonal topic space), and later explicitly distilled as features in the feature construction algorithm.

4. Model generation

In this section, we apply the EM (Expectation Maximization) algorithm to maximize the likelihood function in Eq. (8).

4.1. EM derivation

Similar to the PLSA model, there are two steps in our model generation stage: (i) in the expectation (E) step, we estimate the posterior probabilities for all the latent variables θ' and (ii) in the maximization (M) step, we update the parameters for the posterior probabilities obtained in the previous E-step.

The lower bound (Jensen's inequality) $\mathcal{L}_0$ of Eq. (8) is:

$$\mathcal{L}_0 = \sum_{\theta'} q(\theta')\left\{\alpha \log\frac{P(\mathbb{T},\theta'|\Lambda)}{q(\theta')} + (1-\alpha)\log\frac{P(\mathbb{A},\theta'|\Lambda)}{q(\theta')}\right\}, \qquad (9)$$

where $q(\theta')$ could be an arbitrary function, and here we set $q(\theta') = \alpha P(\theta'|\mathbb{T},\Lambda^{old}) + (1-\alpha)P(\theta'|\mathbb{A},\Lambda^{old})$. Substitute $q(\theta')$ into (9), we have:

$$\begin{aligned}\mathcal{L}_0 &= \sum_{\theta'} q(\theta') \times (\alpha \log(P(\theta'|\mathbb{T},\Lambda)) + (1-\alpha)\log(P(\theta'|\mathbb{A},\Lambda)))\}\\ &\quad \underbrace{-\textstyle\sum_{\theta'} q(\theta') \times (\alpha \log(P(\theta'|\mathbb{T},\Lambda^{old})) + (1-\alpha)\log(P(\theta'|\mathbb{A},\Lambda^{old})))\}}_{const}\\ &= \mathcal{L} + const. \qquad (10)\end{aligned}$$

Therefore, the E step and M step for the solution of Eq. (10) is given as:

E Step:

$$\begin{aligned}P(\theta'|w_j,s_i) &= \frac{P(w_j|\theta')P(\theta'|s_i)}{\sum_{\theta'} P(w_j|\theta')P(\theta'|s_i)},\\ P(\theta'|s_j,s_i) &= \frac{P(s_j|\theta')P(\theta'|s_i)}{\sum_{\theta'} P(s_j|\theta')P(\theta'|s_i)}. \qquad (11)\end{aligned}$$

M Step:

$$\begin{aligned}P(w_j|\theta') &= \sum_i \frac{t_{ij}}{\sum_{j'} t_{ij'}} P(\theta'|w_j,s_i),\\ P(s_j|\theta') &= \sum_i \frac{a_{ij}}{\sum_{j'} a_{ij'}} P(\theta'|s_j,s_i), \qquad (12)\end{aligned}$$

$$\begin{aligned}P(\theta'|s_i) &\propto \alpha \cdot \sum_j \frac{t_{ij}}{\sum_{j'} t_{ij'}} P(\theta'|w_j,s_i)\\ &\quad + (1-\alpha)\cdot \sum_j \frac{a_{ij}}{\sum_{j'} a_{ij'}} P(\theta'|s_j,s_i). \qquad (13)\end{aligned}$$

As we can see from the above equations, the time complexity of model generation phase is $O(I \cdot T \cdot N)$, where I is the number of iterations, T is the number of latent topics and N is the number of total term-document co-occurrences.

4.2. Injecting query

To describe the topic about the query θ_q in the generative model, we build a unigram language model $\{P(w_i|\theta_q)\}$, $w_i \in Narr$, where *Narr* denotes the query narrative for the documents, and w_i denotes the *i*th word in *Narr*. For instance, a language model for the narrative "What are the benefits of drug legalization?" may be represented as (with all the stop words removed):

$$\begin{aligned}&p(\text{benefits}|\theta_q) = 1/3;\\ &p(\text{drug}|\theta_q) = 1/3;\\ &p(\text{legalization}|\theta_q) = 1/3.\end{aligned}$$

We treat the query title and narrative as a pseudo sentence. In the initial phase, the topic about the query θ_q only includes the words in topic title and narrative, when the training process is finished, the word cluster for θ_q is expanded and the relevant words and sentences of query are captured by θ_q.

5. Feature construction and topic balance imposition

To generate summaries that wisely preserve the information in the documents as much as possible while bias the given query, we construct features both from the document perspective and from the query perspective, which are listed in Table 2. Besides, other parameters in algorithms such as β, γ and w are also presented. Specifically, the greedy algorithm for topic balance is given as Algorithms 2 and 3, while the features are distilled as Algorithm 1.

Algorithm 1. The Feature Construction Algorithm

```
Input:
    $p(\theta'|s_i)$ of each sentence, $p(w_{ij})$ of each word, term-frequency matrix $\mathbb{T}$ and similarity matrix $\mathbb{A}$.
Output:
    DN, QN, DR, QR for each sentence.
for each $s_i$ do
    $\mathrm{DR}_i \leftarrow \sum_j a_{i,j}$
    $\mathrm{QR}_i \leftarrow p(\theta_q|s_i) \cdot a(s_i, \mathcal{Q})$
    $\mathrm{DN}_i \leftarrow \sum_j - p(w_{ij}) \cdot \log(p(w_{ij}))$
    $\mathrm{QN}_i \leftarrow (1 - p(\theta_q|s_i)) \cdot \sum_{\theta_j \in \theta} - p(\theta_j|s_i) \cdot \log(p(\theta_j|s_i))$
end for
```

In Algorithm 1, the distilled features are discussed in the follows, and DR_i is calculated as:

Table 2
Parameters and constructed features.

Symbol	Description
DR	Relevance from the document perspective
QR	Relevance from the query perspective
DN	Novelty from the document perspective
QN	Novelty from the query perspective
TB	The feature for topic balance
QB	The feature for query balance
β	The weight of balance
γ	The expected ratio of query
w	The weight of penalty
RA_θ	The ratio of θ in the current summary
$\mathfrak{K}$	The number of sentences in the final summary

$$\text{DR}_i \leftarrow \sum_j a_{i,j}. \tag{14}$$

This feature sums up the similarities between s_i and all the other sentences, which strives to measure the sentence relevance of s_i to the whole document. A large value of DR means that the content in s_i is very similar to other sentences, thus indicates that the information in s_i is mentioned frequently by other sentences in the whole document, i.e., s_i is highly relevant to the document content.

QR$_i$ is measured as:

$$\text{QR}_i \leftarrow p(\theta_q|s_i) \cdot a(s_i, \mathcal{Q}), \tag{15}$$

where $p(\theta_q|s_i)$ is the dependent query relevance in the latent topic space and $a(s_i, \mathcal{Q})$ denotes the direct query relevance (similarity). This feature multiplies the explicit query relevance by the internal query relevance to comprehensively compute the relevance between sentence and the query.

DN$_i$ is obtained as:

$$\text{DN}_i \leftarrow \sum_j - p(w_{ij}) \cdot \log(p(w_{ij})), \tag{16}$$

where w_{ij} is the jth word in the ith sentence, and $p(w_{ij})$ denotes the probability of w_{ij} in the ***document***. DN calculates the information novelty in an entropy-like manner so as to measure the novel information richness in s_i. The rationale for the entropy-like way to compute novelty is detailed as follows. As novelty is the desire for novel information, which could only be concealed in unknown text. Since entropy [31] describes information uncertainty, it estimates the expectation amount of unknown information that could possibly be novel. Thereafter, we consider the entropy-like way somewhat an attempt to discover unknown message (novelty). As is mentioned previously, DN utilizes document probability, which considers the word information richness from the document perspective.

QN$_i$ is computed as:

$$\text{QN}_i \leftarrow (1 - p(\theta_q|s_i)) \cdot \sum_{\theta_j \in \theta} - p(\theta_j|s_i) \cdot \log(p(\theta_j|s_i)). \tag{17}$$

QN$_i$ computes the query novelty of s_i in the latent topic space. Since QN aims to choose novel sentences against the given query, this feature is scaled by $1 - p(\theta_q|s_i)$ for a straightforward query novelty acquisition. Taken further, utilizing the probability vector $p(\theta|s_i)$, this feature computes the novelty in an entropy-like manner to obtain the information richness for s_i among the background latent topics (i.e., novelty against the query). Incidentally, for both DN and QN, before computing novelty in the entropy-like manner, we normalize the probabilities.

As can be seen, the time complexity of Algorithm 1 is $O(S^2)$, where S is the total number of sentences. As the final summary should be biased to the given query, we define an expected percentage γ for the query topic. Suppose the ratio of query is RA_{θ_q}, QB is dynamically updated as Algorithm 3. QB = $\beta(\beta > 1)$ holds for sentences that belong to θ_q when the summary is not so "query-relevant" (query ratio $\leqslant \gamma$). However, when θ_q is excessively emphasized (query ratio > γ), QB is set as $1/\beta$ for query penalty. While for the sentences belonging to other topics, QB remains as 1. As can be seen, the time complexity of the Query Balance is $O(S)$. Incidentally, we categorize sentences into different topics as:

$$\theta^* = \arg\max_{\theta'} p(\theta'|s_i). \tag{18}$$

Algorithm 2. The Greedy algorithm for Topic Balance

```
Input:
  The probabilities $p(s_{i=1\ldots n}|\theta')$, $p(\theta'|s_{i=1\ldots n})$, $p(\theta')$, $\beta$, $\gamma$, the
  features QR, DR, QN, DN and the summary sentence
  number $\mathfrak{K}$.
Output:
  The final summary $\mathfrak{S} = \{x_1, \ldots, x_{\mathfrak{K}}\}$.
Iter ← 1
while $Iter <= \mathfrak{K}$ do
  Set the vector $RA_{\theta'} := 0$
  for each $x_j$ in current summary $\mathfrak{S}$ do
    Find the topic $\theta^*$ of $x_j$ and set $RA_{\theta^*} \leftarrow RA_{\theta^*} + 1$
  end for
  Normalize the vector $RA_{\theta'}$
  $\theta_k \leftarrow \arg\min_{\theta'} RA_{\theta'}$
  Generate a random number $acc$ from $(0,1]$
  if $acc \leqslant p(\theta_k)$ then
    TB ← $\beta$ for each sentence belonging to $\theta_k$
    TB ← 1 for each sentence not belonging to $\theta_k$
  else
    TB ← 1 for each sentence $s_i$
  end if
  QueryBlance($RA_{\theta_q}, \gamma, \beta$)
  for each $s_i$ do
    $R_i \leftarrow \text{DR}_i \cdot \text{QR}_i \cdot \text{DN}_i \cdot \text{QN}_i \cdot \text{TB}_i \cdot \text{QB}_i$
  end for
  for each $s_i$ NOT IN $\mathfrak{S}$ do
    for each $x_j$ IN $\mathfrak{S}$, get its corresponding topic $\theta^*$ do
      Update $R_i = R_i - w \cdot p(s_i|\theta^*) \cdot R_{x_j}$
    end for
  end for
  $x_{Iter} \leftarrow \text{argmax}_{s_i} R_i, \mathcal{S} \leftarrow \mathcal{S} \bigcup x_{Iter}$
  $Iter \leftarrow Iter + 1$
end while
```

As we know, coverage demands the summary to cover different topics reasonably, i.e., the summarization process should balance topic distribution during sentence selection. In this study, QB and TB are proposed to balance the ratios of query and content topics and try to maintain the analogous coverage of different topics in the summary as in the original document. Finally, the algorithm for topic balance is given as Algorithm 2. In the algorithm, a random number is generated to determine whether to balance topic or not. Since topic ratio could be recognized as probability distribution among documents, our idea of topic balance is inspired by the Monte Carlo sampling method [25] for topic ratio adjustment. Similar to QB, TB is updated dynamically to bias specific content topics as the topic ratios change. Besides, it is worth mentioning that we perform topic level redundancy reduction in the sentence selection stage and the time complexity of topic balance is $O(\mathfrak{K} \cdot (3S + S \cdot \mathfrak{K}/2))$.

Algorithm 3. The Query Balance Algorithm QueryBalance($RA_{\theta_q}, \gamma, \beta$)

```
Input:
  The ratio of $RA_{\theta_q}$, $\beta$ and $\gamma$.
Output:
  The feature QB for each sentence.
if $RA_{\theta_q} \leqslant \gamma$ then
  for each $s_i$ belonging to $\theta_q$ then
    QB ← $\beta$
  end for
else
  for each $s_i$ belonging to $\theta_q$ then
    QB ← 1
  end for
end if
```

6. Experiments and analysis

6.1. Data set preparation

We use the benchmark data sets DUC[1]2005 and DUC2006 to testify the effectiveness of our model. Note that each summary task in the dataset is accompanied with a query. In our implementations, each document and query are decomposed into sentences, with all the stop words excluded. Moreover, each word is stemmed using Porter's stemming.[2] Hence, each query and sentence could be represented as a vector of vocabulary, and the word frequencies and sentence similarities are calculated for model generation. Note that we only adopt the statistical information of words and sentences, there is no other tagging or lexical tools in our implementation.

Specifically, the similarities between sentences are computed after stop word elimination and word stemming. The features for sentence ranking are constructed after the generation of probabilistic model (PRCN), where the probabilities for feature quantification are obtained. According to the selected sentence sequence from Algorithm 2, we select the corresponding sentences from the original document as the final summary.

6.2. Evaluation metrics

The popular ROUGE [23] toolkit is adopted for evaluation, which automatically identifies n-grams and stems (Porter's stemming) each word for summary evaluation. And the parameter setting for ROUGE is the same as the official parameter setting.[3] Specifically, we present the details of ROUGE metrics as follows:

ROUGE-N-R is an n-gram recall metric formulated as:

$$\text{ROUGE-N-R} = \frac{\sum_{y \in y} \sum_{gram_n \in y} \text{Count}_{match}(\text{gram}_n)}{\sum_{y \in y} \sum_{gram_n \in y} \text{Count}_{ground}(\text{gram}_n)}. \tag{19}$$

ROUGE-N-P is an n-gram precision metric formulated as:

$$\text{ROUGE-N-P} = \frac{\sum_{y \in \bar{y}} \sum_{gram_n \in y} \text{Count}_{match}(\text{gram}_n)}{\sum_{y \in \bar{y}} \sum_{gram_n \in y} \text{Count}_{pred}(\text{gram}_n)}. \tag{20}$$

ROUGE-N-F is an n-gram F_1 metric formulated as:

$$\text{ROUGE-N-F} = \frac{2\text{ROUGE-N-R} \cdot \text{ROUGE-N-P}}{\text{ROUGE-N-R} + \text{ROUGE-N-P}}, \tag{21}$$

where n denotes the length of the n-gram, and $gram_n \in y$ denotes the n-grams in the ground truth summary y. $\text{Count}_{match}(\text{gram})_n$ is the number of gram_n co-occurring in the predicted summary $\bar{y}$ and the ground truth summary y, $\text{Count}_{ground}(\text{gram})_n$ is the occurrence number of gram_n in the ground truth summary y, and $\text{Count}_{pred}(\text{gram}_n)$ denotes the occurrence number of gram_n in the predicted summary $\bar{y}$. Besides, we also introduce the ROUGE-W metric that computes the ROUGE scores due to the weighted longest common subsequence.

6.3. Results and evaluation

6.3.1. Overall performance comparison

In our implementation, the parameters α, β, γ and w are set empirically, which would be investigated and discussed in detail later. Besides, for summary generation, we discard sentences that are too long or too short. Specifically, we only choose sentences whose length is between 10 and 35 words.

To the best of our knowledge, the method proposed by [21] is the only algorithm that explicitly considers coverage, novelty and balance in query-focused multi-document summarization. We compare our results with those in their model, meanwhile, we also compare our results to some other baseline methods for a more comprehensive evaluation, which are listed in Table 3. As annotated in the table, NIST denotes the NIST baseline, Random denotes random sentence selection, LDA is the method in [2], which only models document word co-occurrences. NA denotes NCBsum-A while NK represents NCBsum-K [21], the MultiMR and manifold ranking are from [39].

The results for PRCN are reported in Table 3, where parameters are empirically tuned as $\alpha = 0.5$, $\beta = 2.8$, $\gamma = 0.4$ and $w = 0.8$. As shown in the table, our model outperforms other methods in ROUGE-2 and ROUGE-W measurements, which strongly testifies the effectiveness of PRCN, i.e., our method wisely considers relevance, coverage, and novelty with respect to query and document content so as to benefit summary generation.

Since the NIST baseline and the Random method do not take account of relevance, coverage or novelty, their performance are in the bottom. LDA is implemented aiming to cover the main topics of documents, yet without consideration of relevance nor novelty, it slightly outperforms the NIST and Random. Although the Manifold and MultiMR carefully consider biased information richness

Table 3
F-measure results. For Rouge-1-F, NA outperforms PRCN 4.12% better, with 95% confidence interval [−0.08 , 0.04], while PRCN outperforms NA on Rouge-2-F 2.48% better with 95% confidence interval [−0.03, 0.04], and on Rouge-W-F 1.48% better with 95% confidence interval [−0.02, 0.02].

	Rouge-1-F	Rouge-2-F	Rouge-W-F
DUC2005			
NIST	0.28760	0.04195	0.09874
Random	0.30994	0.03892	0.10616
LDA	0.28952	0.04278	0.11100
Manifold	0.37493	0.07410	0.12916
MultiMR	0.36978	0.06790	0.12878
NA	**0.38868**	0.07878	0.13424
NK	0.38570	0.07678	0.13282
PRCN	0.37282	**0.08070**	**0.13535**
DUC2006			
NIST	0.32095	0.05269	0.10993
Random	0.34421	0.05133	0.11779
LDA	0.31697	0.05329	0.11849
Manifold	0.38813	0.08168	0.13396
MultiMR	0.40189	0.08441	0.13943
NA	**0.40869**	0.08981	0.14074
NK	0.40515	0.08698	0.13972
PRCN	0.39254	**0.0922**	**0.14372**

[1] http://www-nlpir.nist.gov/projects/duc/data.html.

[2] http://www.tartarus.org/martin/PorterStemmer/.

[3] -n 4 -w 1.2 -m 2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -d.

(relevance and coverage to some extent) and information novelty, there is no significant effort paid to topic balance (or say appropriate topic coverage distribution assignment). As can be seen, the outperforming methods are NA, NK and PRCN that incorporate relevance, balanced coverage, and novelty. However, our model differs from NA and NK in that our model constructs features with consideration of both user query and document content, while NA and NK only take account of novelty, coverage, and balance from the query-aware angle.

6.3.2. Parameter tuning

In our model, α is used to tune the tradeoff between relevance and coverage modeling, β is utilized to adjust topic balance, γ is adopted as the expected query percentage, and w penalizes topic redundancy. We carry out systematic experiments with different α, β, γ and w to see their influence.

Different α (Eq. (8)) differently weighs coverage modeling to relevance modeling. We assign α to values ranging from 0 to 1 to investigate the influence of α, with β, γ and w fixed. Table 4 illustrates the experimental results. For DUC2005 and DUC2006, the best performance is achieved when $\alpha = 0.5$. Moreover, when $\alpha = 0$ (no coverage modeling) or $\alpha = 1$ (no relevance modeling), the performances are far from the best. The reason is that, both relevance and coverage modeling are critical for summarization, overemphasis on either criterion will result a imperfect summary. As can be inferred, the balance between relevance and coverage that α brings in our framework benefits our summarization task.

β (Algorithms 2 and 3) denotes the strength to bias sentences belonging to specific topics whose ratios need to increase. We conduct systematic experiments with different β ranging from 1 to 5 to evaluate the influence of β, while α, γ and w are fixed. Table 5 illustrates the experimental results and we can see that for DUC2005 and DUC2006, the best performance is achieved when $\beta = 2.8$. Actually, the worst performance is obtained when $\beta = 1$, i.e., no topic balance is imposed. This strongly supports our idea of topic balance to specific topics so as to maintain moderate topic ratio (coverage). Further, when β increases, the ROUGE scores decrease, this suggest that overemphasis on topic coverage would lead to imperfect summaries. The rationale behind is: If β is overemphasized, the sentence selection mechanism would only select sentences to maintain appropriate topic coverage, which weakens the influences of relevance and novelty.

Table 5
Parameter investigation: ROUGE scores VS β. For Rouge-1-F, beta=2.8 outperforms other beta settings of 0.16% with 95% interval [−0.07 to 0.07], and for Rouge-2-F, beta=2.8 outperforms other beta settings of 0.75% with 95% interval [−0.04 to 0.03], and Rouge-W-F 0.30% with 95% interval [−0.03 to 0.03].

	Rouge-1-F	Rouge-2-F	Rouge-W-F
DUC2005 ($\alpha = 0.4$, $\gamma = 0.5$, $w = 0.8$)			
$\beta = 1$	0.36915	0.08038	0.13487
$\beta = 1.4$	0.36725	0.07870	0.13397
$\beta = 1.8$	0.36699	0.07942	0.13416
$\beta = 2.2$	0.36736	0.07950	0.13422
$\beta = 2.5$	0.36873	0.08007	0.13469
$\boldsymbol{\beta = 2.8}$	**0.36933**	**0.08052**	**0.13508**
$\beta = 3$	0.36885	0.08040	0.13472
$\beta = 3.5$	0.36920	0.08044	0.13481
$\beta = 4$	0.36928	0.08031	0.13472
$\beta = 5$	0.36926	0.08036	0.13479
DUC2006 ($\alpha = 0.4$, $\gamma = 0.5$, $w = 0.8$)			
$\beta = 1$	0.39222	0.09066	0.14294
$\beta = 1.4$	0.39271	0.09069	0.143059
$\beta = 1.8$	0.39260	0.09120	0.14332
$\beta = 2.2$	0.39145	0.09148	0.14318
$\beta = 2.5$	0.39196	0.09186	0.14354
$\boldsymbol{\beta = 2.8}$	**0.39276**	**0.09220**	**0.14375**
$\beta = 3$	0.39256	0.09181	0.14362
$\beta = 3.5$	0.39285	0.09203	0.14362
$\beta = 4$	0.39248	0.09170	0.14360
$\beta = 5$	0.39244	0.09165	0.14360

γ denotes the expected percentage of the topic describing the query. We implement a set of experiments with different γ ranging from 0 to 1 to evaluate the influence of γ, while α, β and w are fixed. Table 6 illustrates the experimental results and we can see that the best performance is achieved when $\gamma = 0.4$. γ is utilized to quantify the extent of query preference. As can be seen from the table, when

Table 4
Parameter investigation: ROUGE scores VS α. For Rouge-1-F, alpha=0.5 outperforms other alpha settings of 8.12% with 95% interval [−0.04 to −0.01], and for Rouge-2-F, alpha=0.5 outperforms other alpha settings of 12.4% with 95% interval [−0.02 to −0.00], and Rouge-W-F of 7.4% with 95% interval [−0.01 to −0.00].

	Rouge-1-F	Rouge-2-F	Rouge-W-F
DUC2005 ($\beta = 1.8$, $\gamma = 0.4$, $w = 0.8$)			
$\alpha = 0$	0.34671	0.07244	0.12540
$\alpha = 0.1$	0.36610	0.07615	0.13301
$\alpha = 0.2$	0.36722	0.07724	0.13382
$\alpha = 0.3$	0.36613	0.07861	0.13344
$\alpha = 0.4$	0.36792	0.07997	0.13464
$\boldsymbol{\alpha = 0.5}$	**0.36930**	**0.08061**	**0.13494**
$\alpha = 0.6$	0.36568	0.07866	0.13341
$\alpha = 0.7$	0.36848	0.07868	0.13407
$\alpha = 0.8$	0.36812	0.07826	0.13408
$\alpha = 0.9$	0.36766	0.07787	0.13396
$\alpha = 1$	0.36791	0.07726	0.13410
DUC2006 ($\beta = 1.8$, $\gamma = 0.4$, $w = 0.8$)			
$\alpha = 0$	0.38768	0.08810	0.14093
$\alpha = 0.1$	0.39237	0.08896	0.14234
$\alpha = 0.2$	0.39246	0.09023	0.14200
$\alpha = 0.3$	0.39078	0.08892	0.14172
$\alpha = 0.4$	0.39196	0.09186	0.14354
$\boldsymbol{\alpha = 0.5}$	**0.39378**	**0.09211**	**0.14367**
$\alpha = 0.6$	0.38729	0.08590	0.14108
$\alpha = 0.7$	0.39331	0.08820	0.14312
$\alpha = 0.8$	0.39128	0.08851	0.14379
$\alpha = 0.9$	0.39098	0.08800	0.14368
$\alpha = 1$	0.39033	0.08720	0.14302

Table 6
Parameter investigation: ROUGE scores VS γ. For Rouge-1-F, gamma=0.4 outperforms other beta settings of 1.1% with 95% interval [−0.07 to 0.06], and for Rouge-2-F, gamma=0.4 outperforms other gamma settings of 3.7% with 95% interval [−0.04 to 0.03], and Rouge-W-F 0.7% with 95% interval [−0.03 to 0.03].

	Rouge-1-F	Rouge-2-F	Rouge-W-F
DUC2005 ($\alpha = 0.5$, $\beta = 2.2$, $w = 0.8$)			
$\gamma = 0$	0.36810	0.07876	0.13447
$\gamma = 0.1$	0.36810	0.07876	0.13447
$\gamma = 0.2$	0.36810	0.07876	0.13447
$\gamma = 0.3$	0.36865	0.07929	0.13488
$\boldsymbol{\gamma = 0.4}$	**0.37278**	**0.08056**	**0.13523**
$\gamma = 0.5$	0.36759	0.07912	0.13395
$\gamma = 0.6$	0.36790	0.07883	0.13394
$\gamma = 0.7$	0.36751	0.07821	0.13397
$\gamma = 0.8$	0.36827	0.07782	0.13425
$\gamma = 0.9$	0.36818	0.07770	0.13419
$\gamma = 1$	0.36726	0.07696	0.13352
DUC2006 ($\alpha = 0.5$, $\beta = 2.2$, $w = 0.8$)			
$\gamma = 0$	0.38761	0.08799	0.14083
$\gamma = 0.1$	0.39208	0.09095	0.14342
$\gamma = 0.2$	0.39212	0.09162	0.14322
$\gamma = 0.3$	0.39192	0.09080	0.14312
$\boldsymbol{\gamma = 0.4}$	**0.39354**	**0.09212**	**0.14411**
$\gamma = 0.5$	0.38821	0.08731	0.14234
$\gamma = 0.6$	0.38706	0.08565	0.14094
$\gamma = 0.7$	0.39323	0.09102	0.14382
$\gamma = 0.8$	0.39309	0.08835	0.14381
$\gamma = 0.9$	0.39292	0.08774	0.14360
$\gamma = 1$	0.39208	0.09095	0.14342

Table 7
Parameter investigation: ROUGE scores VS *w*. For Rouge-1-F, w=0.8 outperforms other beta settings of 0.05% with 95% interval [−0.07 to 0.06], and for Rouge-2-F, w=0.8 outperforms other w settings of 0.25% with 95% interval [−0.04 to 0.03], and Rouge-W-F 0.22% with 95% interval [−0.03 to 0.03].

	Rouge-1-F	Rouge-2-F	Rouge-W-F
DUC2005 ($\alpha = 0.5$, $\beta = 2.8$, $\gamma = 0.4$)			
$w = 0$	0.36908	0.08037	0.13491
$w = 0.1$	0.36904	0.08044	0.13490
$w = 0.2$	0.36907	0.08048	0.13485
$w = 0.3$	0.36903	0.08039	0.13482
$w = 0.4$	0.36880	0.08030	0.13475
$w = 0.5$	0.36918	0.08059	0.13492
$w = 0.6$	0.36930	0.08061	0.13494
$w = 0.7$	0.36930	0.08061	0.13494
$w = 0.8$	**0.37282**	**0.08070**	**0.13535**
$w = 0.9$	0.36935	0.08060	0.13490
$w = 1$	0.36935	0.08060	0.13490
DUC2006 ($\alpha = 0.5$, $\beta = 2.8$, $\gamma = 0.4$)			
$w = 0$	0.39252	0.09207	0.14368
$w = 0.1$	0.39252	0.09207	0.14368
$w = 0.2$	0.39247	0.09208	0.14367
$w = 0.3$	0.39247	0.09208	0.14367
$w = 0.4$	0.39247	0.09208	0.14367
$w = 0.5$	0.39247	0.09208	0.14367
$w = 0.6$	0.39247	0.09208	0.14367
$w = 0.7$	0.39247	0.09208	0.14367
$w = 0.8$	**0.39254**	**0.09220**	**0.14372**
$w = 0.9$	0.39236	0.09211	0.14366
$w = 1$	0.39244	0.09211	0.14366

$\gamma = 0$ or $\gamma = 1$ the model gets poor performance, this corroborates our idea of maintaining query relevance to suitable extent. For summarization, there should be neither too little nor too much concern on query relevance.

In Algorithm 2, w is introduced for topic redundancy reduction. We carry out systematic experiments with different w ranging from 0 to 1 to evaluate the influence of w, while α, β and γ are fixed. As can be seen from Table 7, the best performance is achieved when $w = 0.8$. Interestingly, results change little as w varies. The reason is that as we execute topic balance and query balance in our framework, appropriate topic distribution is carefully maintained, thereafter topic redundancy is implicitly eliminated. Since our algorithm wisely weighs each topic, it avoids excessive concentration (i.e., redundancy) on certain topics to some extent.

6.3.3. Feature investigation

In this section, we evaluate the effectiveness of each feature we defined in Section 5. Specifically, in this experimentation, we rank the salience score of each sentence as $R_i \leftarrow \mathrm{QR}_i$, $R_i \leftarrow \mathrm{DR}_i$, $R_i \leftarrow \mathrm{QN}_i$, $R_i \leftarrow \mathrm{DN}_i$, $R_i \leftarrow \mathrm{QB}_i$ and $R_i \leftarrow \mathrm{TB}_i$, respectively. The performance of each feature is separately studied with all the parameters fixed.

Fig. 1 illustrates the performance of each constructed feature on DUC05 and DUC06, where QN denotes the ranking method based on the feature QN, and others are name similarly, while Overall is ranking utilizing all the features. As it is shown in the figures, QR obtains the worst performance while DN and QN perform in the lead, which indicates that novelty plays a more important role than other features in the process of summarization. To further validate this inference, we separately discard features QR, DR, QN, DN, TB and QB, one by one from the ranking equation $R_i \leftarrow \mathrm{DR}_i \cdot \mathrm{QR}_i \cdot \mathrm{DN}_i \cdot \mathrm{QN}_i \cdot \mathrm{TB}_i \cdot \mathrm{QB}_i$ in Algorithm 2.

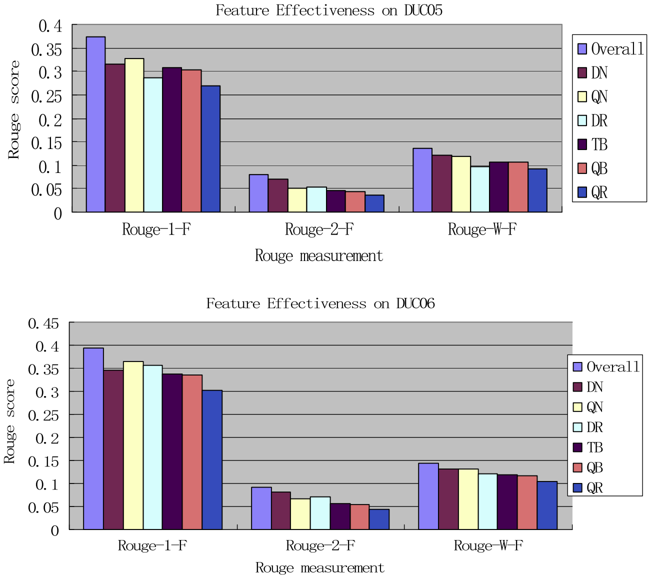

Fig. 1. Comparison of different features: for both DUC05 and DUC06, QN and DN outperforms other features for different ROUGE metrics, which indicates the importance of novelty in query-focused multi-document summarization.

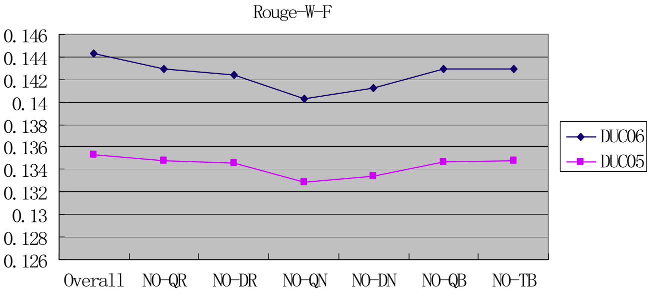

Fig. 2. ROUGE-W-F with different features excluded: for both DUC05 and DUC06, NO-QN and NO-DN achieves the worst performance among all the NO-features, which again testifies the importance of QN and DN in query-focused multi-document summarization.

As shown in Fig. 2, NO-QN represents ranking excluding the feature QN, and others are named similarly. As can be seen, NO-QN and NO-DN acquire the lowest scores for DUC05 and DUC06, which again supports the significance of novelty for summarization. The result that novelty outperforms other features suggests that query-focused multi-document summarization tasks demand more on new information(novelty). Meanwhile, the result that in both Figs. 1 and 2, Overall outperforms all the rest results suggests that each feature benefits the summarization task. This strongly corroborates our idea of incorporating relevance, coverage, and novelty both from the query and the document perspective.

7. Conclusion and future work

In this paper, we go beyond the independent relevance assumption, and propose a Probabilistic-modeling Relevance, Coverage, and Novelty (PRCN) framework to model topic relevance and coverage, where a reference topic model incorporating query is utilized for dependent sentence relevance measurement. Furthermore, our work makes contribution in constructing and quantifying a set of features describing relevance, novelty, and topic balance both from the document and from the query perspective. Besides the constructed features, we also develop a greedy topic balance algorithm for sentence ranking and extraction. Experiments are conducted to verify the effectiveness of our model, and the results show that: (1) Dependently modeling **relevance** and **coverage** under a joint probabilistic framework, PRCN proves to be effective for query-focused multi-document summarization; (2) PRCN could effectively **balance** query (user profile) and content topics with appropriate **coverage** through a greedy algorithm; (3)

Document Relevance/Query Relevance and Document Novelty/Query Novelty are acquired as the constructed sentence features, while Topic Balance/Query Balance are attained in the greedy balance algorithm. Among all the constructed features, Document Novelty/Query Novelty are demonstrated to be more potent than other features in query-focused multi-document summarization.

On the other hand, as is discussed in the paper, LDA-based topic models exhibit a set of merits compared to PLSA-based methods. Our future work would focus on developing models based on LDA for document summarization.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 61175052, 61203297, 60933004, 61035003), National High-tech R&D Program of China (863 Program) (Nos. 2013AA01A606, 2012AA011003), National Program on Key Basic Research Project (973 Program) (No. 2013CB329502).

References

[1] A. Berger, V.O. Mittal, Query-relevant summraization using faqs, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguisitics (ACL'00), 2000. pp. 294–301.

[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.

[3] J. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 26th Annual International ACM SIGIR Conference (SIGIR'98), 1998, pp. 335–336.

[4] H. Chen, D.R. Karger, Less is more probabilistic models for retrieving fewer relevant documents, in: Proceedings of the 34th Annual International ACM SIGIR Conference (SIGIR'06), 2006, pp. 429–436.

[5] C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Bttcher, I. MacKinnon, Novelty and diversity in information retrieval evaluation, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 659–666.

[6] D. Cohn, H. Chang, Learning to probabilistically identify authoritative documents, in: Proceedings of the 17th International Conference on Machine Learning (ICML'00), 2000, pp. 167–174.

[7] D. Cohn, T. Hofmann, in: Proceedings of Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference, 2001, pp. 430–436.

[8] J.M. Conroy, D.P. Oleary, Text summarization via hidden Markov models, in: Proceedings of the 24th Annual International ACM SIGIR Conference (SIGIR'01), 2001, pp. 406–407.

[9] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[10] J. Goldstein, V.O. Mittal, J. Carbonell, M. Kantrowitz, Multi-document summarization by sentence extraction, in: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, 2000, pp. 40–48.

[11] Y. Gong, X. Liu, Generic text summarization using relevance measure and latent semantic analysis, in: Proceedings of the 29th Annual International ACM SIGIR Conference (SIGIR'01), 2001, pp. 19–25.

[12] S. Gupta, A. Nenkova, D. Jurafsky, Measuring importance and query relevance in topic-focused multi-document summarization, in: Proceedings of the 45th Annual Meeting on Association for Computational Linguisitics (ACL'07), 2007, pp. 193–196.

[13] T. Hirao, J. Suzuki, H. Isozaki, E. Maeda, Ntt's multiple document summarization system for duc2003, in: Proceedings of Document Understanding Conference, 2003.

[14] T. Hofmann, Probabilistic latent semantic analysis, in: Proceedings of the 15th Conference on Uncertainty in AI, 1999, pp. 289–296.

[15] S.R. Joty, Automatic annotation techniques for supervised an semi-supervised query-focused summarization, Technical report, The University of British Columbia, source:http://www.cs.ubc.ca/~carenini/TEACHING/CPSC503-10/FINAL-REPORTS-08/ShafiqReport.pdf.

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

[17] J. Kupiec, J. Pedersen, F. Chen, A trainable document summarizer, in: Proceedings of SIGIR'95, 1995, pp. 68–73.

[18] J. Lafferty, C. Zhai, Document language models, query models, and risk minimization for information retrieval, in: Proceedings of the 29th Annual International ACM SIGIR Conference (SIGIR'01), 2001, pp. 111–119.

[19] J. Lafferty, C. Zhai, Probabilistic relevance models based on document and query generation, in: Language Modeling and Information Retrieval, 2002, pp. 1–10.

[20] L. Li, K. Zhou, G.-R. Xue, H. Zha, Y. Yu, Enhancing diversity, coverage and balance for summarization through structure learning, in: Proceedings of the 18th International World Wide Web Conference (WWW'09), 2009, pp. 71–80.

[21] X. Li, Y.-D. Shen, L. Du, C.-Y. Xiong, Exploiting novelty, coverage and balance for topic-focused multi-document summarization, in: Proceedings of the 19th ACM International Conferences on Information and Knowledge Management (CIKM'10), 2010, pp. 1765–1768.

[22] C. Lin, Y. He, Joint sentiment/topic model for sentiment analysis, in: Proc. of CIKM'09, 2011, pp. 132–141.

[23] C.-Y. Lin, E. Hovy, Automatic evaluation of summaries using *n*-gram co-occurrence statics, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03), 2003, pp. 71–78.

[24] H. Luhn, The automatic creation of literature abstracts, IBM Journal of research and development 2 (2) (1958) 159–165.

[25] N. Metropolis, The beginning of the Monte Carlo method. Los Alamos Science 15 (584) (1987) 125–130.

[26] D. Metzler, T. Kanungo, F. Chen, Machine learned sentence selection strategies for query-biased summarization, in: Workshop of SIGIR'08, 2008.

[27] J.L. Neto, A.A. Freitas, C, A.A.Kaestner, Automatic text summarization using a machine learning approach, in: Proceedings of the 16th Brazilian Symposium on Artificial Intelligence, 2002, pp. 205–215.

[28] A.Y. Ng, A.X. Zheng, M.I. Jordan, Link analysis, eigenvectors and stability, in: Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI'01), 2001, pp. 903–910.

[29] T. Nomoto, A new approach to unsupervised text summarization, in: Proceedings of the 29th Annual International ACM SIGIR Conference (SIGIR'01), 2001, pp. 26–34.

[30] S.E. Robertson, The probability ranking principle in ir, Journal of Documentation 33 (4) (1977) 294–304.

[31] C.E. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mobile Computing and Communications Review 5 (1) (2001) 3–55.

[32] C. Shen, T. Li, Multi-document summarization via the minimum dominating set, in Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10), 2007, pp. 2903–2908.

[33] D. Shen, J.-T. Sun, H. Li, Q. Yang, Z. Chen, Document summarization using conditional random fields, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), 2007, pp. 2862–2867.

[34] A. Tombros, C. van Rijsbergen, Query-sensitive similarity measures for information retrieval, Knowledge and Information Systems 6 (2004) 617–642.

[35] J. Ulrich, G. Carenini, G. Murray, R. Ng, Regression based summarization of email conversations, in: Proceedings of 3rd Int'l AAAI Conference on Weblogs and Social Media (ICWSM'09), 2009.

[36] E.M. Voorhees, Variations in relevance judgements and the measurement of retrieval effectiveness, in: Proceedings of the 26th Annual International ACM SIGIR Conference (SIGIR'98), 1998, pp. 315–323.

[37] X. Wan, J. Xiao, Graph-based multi-modality learning for topic-focused multi-document summarization, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09), 2009, pp. 1586–1591.

[38] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: Proceedings of the 31st Annual International ACM SIGIR Conference (SIGIR'08), 2008, pp. 299–306.

[39] X. Wan, J. Yang, J. Xiao, Manifold-ranking based topic-focused multi-document summarization, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), 2007, pp. 2903–2908.

[40] D. Wang, S. Zhu, T. Li, Y. Gong, Multi-document summarization using sentence-based topic models, in: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (the ACL-IJCNLP'09), 2009, pp. 297–300.

[41] F. Wei, W. Li, Q. Lu, Y. He, Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization, in: Proceedings of the 36th Annual International ACM SIGIR Conference (SIGIR'08), 2008, pp. 283–290.

[42] Y. Xu, H. Yin, Novelty and topicality in interactive information retrieval, Journal of the American Society for Information Science and Technology 59 (2) (2008) 201–215.

[43] H. Zha, Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering, in: Proceedings of the 30th Annual International ACM SIGIR Conference (SIGIR'02), 2002, pp. 26–34.

[44] C. Zhai, W.W. Cohen, J. Lafferty, Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, in: Proceedings of the 31st Annual International ACM SIGIR Conference (SIGIR'03), 2003, pp. 10–17.

[45] Y. Zhang, J. Callan, T. Minka, Novelty and redundancy detection in adaptive filtering, in: Proceedings of the 30th Annual International ACM SIGIR Conference (SIGIR'02), 2002, pp. 81–88.

[46] Z. Zhou, H. Dai, Query-sensitive similarity measure for content-based image retrieval, in: Proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06), 2006, pp. 1211–1215.