Heterogeneous-Length Text Topic Modeling for Reader-Aware Multi-Document Summarization

JIPENG QIANG, Yangzhou University
PING CHEN, WEI DING, and TONG WANG, University of Massachusetts Boston
FEI XIE, Hefei Normal University
XINDONG WU, Mininglamp Academy of Sciences, Minininglamp and Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education

More and more user comments like Tweets are available, which often contain user concerns. In order to meet the demands of users, a good summary generating from multiple documents should consider reader interests as reflected in reader comments. In this article, we focus on how to generate a summary from multi-document documents by considering reader comments, named as reader-aware multi-document summarization (RA-MDS). We present an innovative topic-based method for RA-MDA, which exploits latent topics to obtain the most salient and lessen redundancy summary from multiple documents. Since finding latent topics for RA-MDS is a crucial step, we also present a Heterogeneous-length Text Topic Modeling (HTTM) to extract topics from the corpus that includes both news reports and user comments, denoted as heterogeneous-length texts. In this case, the latent topics extract by HTTM cover not only important aspects of the event, but also aspects that attract reader interests. Comparisons on summary benchmark datasets also confirm that the proposed RA-MDS method is effective in improving the quality of extracted summaries. In addition, experimental results demonstrate that the proposed topic modeling method outperforms existing topic modeling algorithms.

CCS Concepts: • **Information systems** → *Data mining*; *Document representation*; • **Applied computing** → *Document management and text processing*;

Additional Key Words and Phrases: Topic modeling, LDA, heterogeneous-length text, multi-document summarization

This research is partially supported by the National Key Research and Development Program of China under grant 2016YFB1000900; the National Natural Science Foundation of China under grant 61703362; the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education, China, under grant IRT17R32; the Natural Science Foundation of Jiangsu Province of China under grant BK20170513; and the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province of China under grant 17KJB520045.
Authors' addresses: J. Qiang (Corresponding author), Department of Computer Science, Yangzhou University, China; email: jpqiang@yzu.edu.cn; P. Chen, W. Ding, and T. Wang, Department of Computer Science, University of Massachusetts Boston; emails: {ping.chen, wei.ding, tong.wang}@umb.edu; F. Xie, Department of Computer Science and Technology, Hefei Normal University, China; email: xiefeihf@gmail.com; X. Wu, Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Ministry of Education, Hefei, China and Mininglamp Academy of Sciences, Minininglamp, Beijing, China, 100084; email: wuxindong@mininglamp.com.

1556-4681/2019/08-ART42 $15.00
https://doi.org/10.1145/3333030

ACM Reference format:
Jipeng Qiang, Ping Chen, Wei Ding, Tong Wang, Fei Xie, and Xindong Wu. 2019. Heterogeneous-Length Text Topic Modeling for Reader-Aware Multi-Document Summarization. *ACM Trans. Knowl. Discov. Data* 13, 4, Article 42 (August 2019), 21 pages.
https://doi.org/10.1145/3333030

1 INTRODUCTION

Multi-document summarization (MDS) has attracted much attention in recent years, which can help to generate a summary from a set of documents about a specific event [1, 2]. The aim of MDS is to extract a succinct summary from multiple documents by reducing information redundancy. Most of the existing summarization systems focus on generating a summary from long texts, e.g., news reports. In recent years, the proliferation of social media, comments, and discussion groups have ushered in an era of information overload. Different from traditional documents, texts from social media or comments are very short, denoted as short texts [3, 4]. Suppose we generate a summary from news reports, it is unreasonable to discount these user comments when extracting a summary from news reports. For example, one hot event in 2018 was "G7 summit." After the outbreak of this event, lots of reposts were posted on different news medias in Figure 1. Most existing summarization systems generated summaries from news reports ignoring reader comments. From Figure 1, we can see that some information from social media (e.g., Twitter) were interested by many readers. Therefore, if MDS can jointly consider long texts and short texts when generating the summary, the generated summary can cover not only important aspects of long texts, but also aspects that attract reader interests as reflected in short texts. The problem is named as reader-aware multi-document summarization (RA-MDS).

The problem has the following two challenges. One challenge is how to conduct salience calculation by jointly considering the focus of news reports and reader interests revealed by comments. The other challenge is that reader comments from social media are very noisy, grammatically and informatively. To solve this problem, only Li et al. [5] proposed a sparse-coding-based method by jointly considering news reports and reader comments. To tackle the above challenges, we try to use topic-based methods for RA-MDS to capture the events being covered by news reports and reader comments. Here, the corpus includes both short texts and long texts, referred to as heterogeneous-length texts. For topic-based summarization methods, finding good latent topics is a crucial step. The existing topic-based summarization methods for MDS [6] often adopt Latent Dirichlet Allocation (LDA) [7] and its variations [7, 8]. These methods are also based on this complex assumption that a text contains multiple topics, which is usually fit for long texts. But for short text, the simple assumption that each text only has one topic achieves better results [9]. For heterogeneous-length texts, both the complex assumption and the simple assumption cannot generate satisfying results.

Therefore, for generating a good summary, we first propose a novel Heterogeneous-length Text Topic Modeling (HTTM). HTTM extracts topics from heterogeneous-length texts based on both the complex assumption and the simple assumption. After mining the topics from heterogeneous-length texts, we present a RA-MDS method to capture the events being covered by news reports and comments, and form the summary from the sentences from news reports. The summary should contain not only important aspects of the event, but also aspects that attract reader interests as reflected in reader comments. In addition, our RA-MDS method considers content coverage and non-redundancy. To emphasize an important point, HTTM can be used to different types of heterogeneous-length texts. In this article, for a better explanation of our innovation, we use news/comments as an example to present this article. The key contribution of the article is twofold.

Fig. 1. The information of G7 summit from news websites and Twitter.

(a) *Heterogeneous-length text topic modeling*: For better to discover the latent topics in heterogeneous-length texts, we first present a novel HTTM. Different from existing topic modelings under a single assumption, HTTM samples a topic for each short text and a set of topics for each long text of a collection. We conduct comprehensive experiments to evaluate HTTM and to demonstrate the effectiveness of our model.

(b) *Reader-aware multi-document summarization*: In this article, we utilize latent topics discovered by HTTM from news reports and reader comments to generate summaries. Such a HTTM-based method has the following noticeable advantages: (1) the generated summaries can not only capture the events being covered by the reports, but also consider the content of reader comments; and (2) the approach is purely statistical and do not involve the structure of the documents or of the sentences in terms of grammar and the meanings conveyed by the words. Experimental results in synthetic datasets and real-world datasets show that our approach outperforms the top competitors.

The remainder of the article is organized as follows. In Section 2, we review related work. Section 3 gives MDS method by utilizing the latent topics by HTTM. Section 4 shows experimental results. Finally, Section 5 concludes the article.

2 RELATED WORK

The proposed research is closely related to topic modeling and MDS.

2.1 Topic Modeling

Based on this assumption that each text is modeled over multiple topics, many topic modelings such as LDA [7] and its variations [3, 10] achieved promising results. Along with the emergence and popularity of social communications (e.g., Twitter and Facebook), short text has become an important information source. Inferring topics from the overwhelming amount of short texts becomes a critical but challenging task for many text analysis tasks [11, 12]. Existing traditional methods for long texts such as probabilistic latent semantic analysis (PLSA) [13] and LDA [7] cannot solve this problem very well since only very limited word co-occurrence information is available in short texts.

Compared with the assumption that each text is modeled over multiple topics used in long texts, this assumption that each text only belongs to one topic works well on short texts [9]. Therefore,

many topic modelings adopted this assumption to generate latent topics from short texts. Nigam et al. [14] proposed a mixture of unigram model, which assumes that each text is generated by one topic. The unigram model is an EM-based algorithm for Dirichlet Multinomial Mixture (DMM) model. Except the basic expectation maximization (EM), a number of inference methods have been used to estimate the parameters including variation inference and Gibbs sampling. For example, Yu et al. [15] proposed the DMAFP model based on variational inference algorithm [16]. Yin et al. [9] proposed a collapsed Gibbs sampling algorithm for DMM (abbr. to GSDMM), and found that GSDMM can infer the number of topics automatically. For speeding up the time, Yin et al. proposed a new algorithm as FGSDMM [17] using an online scheme for initialization. Qiang et al. [18] proposed a novel method based on Pitman–Yor Process to capture the power-law phenomenon of the topic distribution.

Some short text topic models try to use the rich global word co-occurrence patterns for inferring latent topics [11, 19]. Due to the adequacy of global word co-occurrences, the sparsity of short texts is mitigated for these models. Biterm topic modeling (BTM) [11] posits that the two words in a biterm share the same topic drawn from a mixture of topics over the whole corpus. WNTM and its variation [19, 20] first construct word co-occurrence network using global word co-occurrences and then infers latent topics from this network, where each word correspond to one node and the weight of each edge stands for the empirical co-occurrence probability of the connected two words.

Self-aggregation based methods are proposed to perform topic modeling and text self-aggregation during topic inference simultaneously. Short texts are merged into long pseudo-documents before topic inference that can help improve word co-occurrence information. This type of methods SATM [21] and PTM [22] posit that each short text is sampled from a long pseudo-document unobserved in current text collection, and infer latent topics from long pseudo-documents, without depending on auxiliary information or metadata.

For topic models for teterogeneous-length texts, Yang et al. [23] proposed an algorithm COTM, which learned topics from both the long documents and the short texts. Different from our article, COTM suppose that there are formal topics and informal topics in short texts.

2.2 Reader-Aware Multi-Document Summarization

Extractive MDS extracts the most informative document components from a set of documents. Extractive summarization is a simple but robust approach without requiring advanced post-processing steps [24–26]. Based on the techniques used in MDS, existing MDS can be divided into the following classifications: term-based methods using term frequency/inverse sentence frequency (TF*ISF) weighting model [27–30], graph-based approaches that produced a similarity graph to help calculate the weight of each sentence [31, 32], and ontology-based approaches that relied on ontology to solve the problems of polysemy and synonymy [2, 33].

Since more and more user generated content is available, Li et al. [5] proposed a new MDS paradigm by incorporating such content regarding the event so as to directly or indirectly improve the generated summaries, called RA-MDS. To deal with this RA-MDS problem, they presented a method based on sparse coding by jointly considering news reports and reader comments. The aim of the article is to extract latent topics from heterogeneous-length texts (short texts and long texts) for generating summaries. Heterogeneous-length texts are just one particular type of heterogeneous sources [3]. We found any one of the above two assumptions may lead to poor inference for heterogeneous-length texts. It is unreasonable to assume that each long text is generated from only one topic, and suppose that each short text is sampled from multiple topics. Motivated by this, we propose a novel HTTM by incorporating the two assumptions together into topic inference. Due to its fundamental nature, HTTM can be used for extracting topics from different types of heterogeneous-length texts. For example, news reports can be as long texts and user comments

Table 1. The Notations of Symbols Used in the Article

D, L, S	Whole corpus, long text set, and short text set
K, V	Number of topics, Size of the vocabulary
d_i^S, d_i^L	ith document of S and L, respectively
$w_{i,j}^L$	jth word in the ith document of L
n_i^L	Number of words the ith document of L
$z_{i,j}^L$	Topic of $w_{i,j}^L$
z_i^S	Topic of d_i^S
n^k	Number of words associated with topic k in D
n^{k,d_i^L}	Number of occurrences of topic k in text d_i^L
$n^{w_{i,j}^L,k}$	Number of occurrences of word $w_{i,j}^L$ belonging to topic k
$\theta_{d_i^L}$	Topic distribution of document d_i^L
ϕ_k	Topic-word multinomial distribution of the kth topic
$\phi_k^{w_p}$	Probability of word w_p belonging to topic ϕ_k
$\theta_{d_i^L}^k$	Probability of topic k belonging to text d_i^L

can be as short texts, Tweets can be as short texts and these Tweets including Web link can be as long texts, and so on. Then, we use the topics mined by HTTM to capture the import content being covered by long texts and short texts, and form the summary using sentences containing the important content.

3 HTTM-BASED SUMMARIZATION

In this section, we discuss how to generate a summary using HTTM, denoted as HTTM-based summarization.

3.1 Overview of HTTM-based Summarization

A document collection D includes a set of long texts $L = \{d_1^L, d_2^L, \ldots, d_i^L, \ldots, d_{|L|}^L\}$ and a set of short texts $S = \{d_1^S, d_2^S, \ldots, d_i^S, \ldots, d_{|S|}^S\}$, where d_l^L represents the ith document of L, d_i^S represents the ith document of S. Here, $|L|$ and $|S|$ is the number of documents in L and S, respectively. After splitting each text into sentences, L can represented as a set of sentences $\{s_1, \ldots, s_m, \ldots, s_M\}$, where s_m represents the mth sentence of L and M is the number of all sentences in L. The notations of symbols used in the article are shown in Table 1.

The aim of this article is to generate a summary $U = \{s_m\}$ ($m \in \{1, \ldots, M\}$) that contains a subset of sentences in L that are representative of L under considering content coverage and non-redundancy. The outline of HTTM-based summarization is shown in Figure 2. Specifically, HTTM-based summarization consists of the following steps:

(1) Obtaining latent topics from a document collection D using HTTM. HTTM utilizes the two assumptions simultaneously, namely the simple assumption that each short text is sampled from one topic and the complex assumption that each long text is generated from multiple topics. Based on the two assumptions, we adopt a collapsed Gibbs sampling to learn the latent topics. Finally, we get the topic distribution for each text (document-topic matrix) and the word distribution for each topic (topic-word matrix).

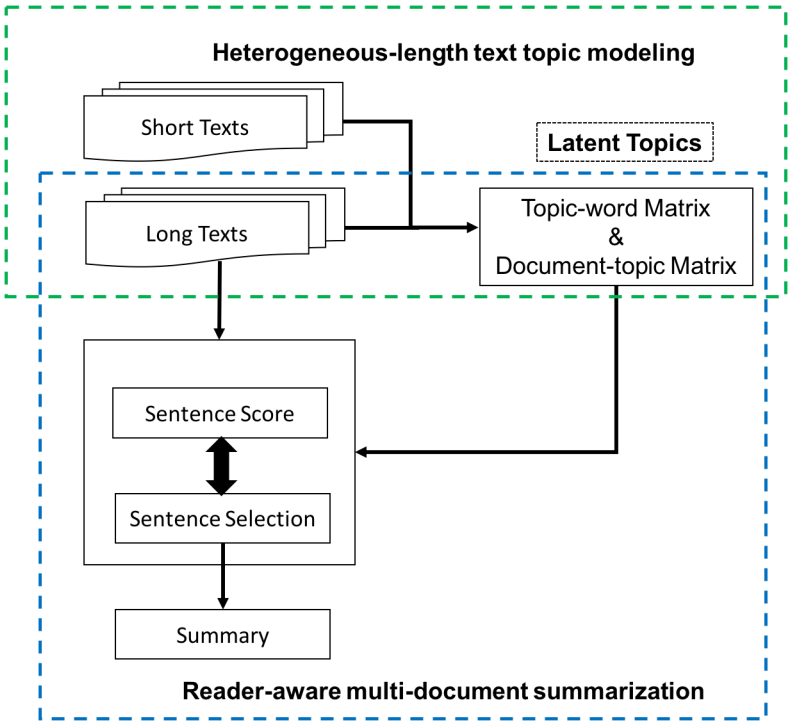


Fig. 2. Flowchart of HTTM-based summarization.

(2) Sentence score. To select the most pertinent and meaningful sentences from L into the summary, sentences in L are ranked according to the previous document-topic matrix and topic-word matrix. Because topic-word matrix contains the important topics covered by short texts and long texts, the scores of the sentences in L are decided by short texts and long texts. Here, we only choose all sentences in L as a source to extract a summary, since the sentences in S are very noisy, grammatically and informatively.
(3) Sentence selection. We iteratively select one sentence that owns the highest weight and is less similar to the previously selected ones, until reaching the length limit.

3.2 Heterogeneous-Length Text Topic Modeling

The generative process of HTTM from heterogeneous-length texts is shown in Figure 3. The left-most portion is used to produce a set of short texts S, which resembles to a mixture of unigrams that each text is inferred from one topic. The rightmost portion in Figure 3 adopts the complex assumption of standard topic models (e.g., LDA) to produce the rest long texts L.

Suppose the corpus contains K clusters. For each cluster k, it draws a word distribution from a Dirichlet distribution with concentration parameter β,

$$\phi_k \sim Dir(\beta). \tag{1}$$

For each long text d_i^L in L, it draws a topic distribution $\theta_{d_i^L}$ from a Dirichlet distribution with concentration parameter α,

$$\theta_{d_i^L} \sim Dir(\alpha). \tag{2}$$

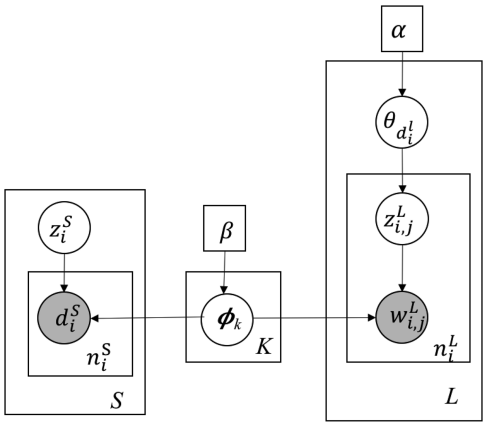


Fig. 3. Graphical representation of the Heterogeneous-length Texts Topics Modeling.

For each word $w_{i,j}^L$ of $d_i^L \in \{1, \ldots, |L|\}$, it draws a topic $z_{i,j}^L$ from the multinomial θ_i^L via,

$$z_{i,j}^L \sim Multinominal(\theta_{d_i^L}). \tag{3}$$

For each word $w_{i,j}^L$ in document d_i^L, it is sampled by the topic-word multinomial distribution $\phi_{z_{i,j}^L}$ via,

$$w_{i,j}^L \sim Multinominal(\phi_{z_{i,j}^L}). \tag{4}$$

For each short text d_i^S in S, it draws a topic z_i^S from $[1, K]$. For each word $w_{i,j}^S$ in document d_i^S, it is sampled by the topic-word multinomial distribution $\phi_{z_i^S}$ via,

$$w_{i,j}^S \sim Multinominal(\phi_{z_i^S}). \tag{5}$$

Given heterogeneous-length texts, the fundamental problem of HTTM is to estimate the posterior distribution of the unseen variables z_i^S and $z_{i,j}^L$ simultaneously, $p(z_i^S, z_{i,j}^L, \phi, \theta \mid D, \alpha, \beta)$. We use collapsed Gibbs sampling to carry out posterior inference for parameter learning under Dirichlet priors. Due to the space limitation, we leave out the details of derivation but give the core formulas in the sampling steps. The hidden multinomial variables text-level variables (z_i^S and $z_{i,j}^L$) are sampled, conditioned on a complete assignment of all other hidden variables. Finally, we update the parameters (θ and ϕ) in a unified framework.

For $w_{i,j}^L$, we use the following conditional probability distribution to infer its topic,

$$p\left(z_{i,j}^L = k \mid D_{-j}, \alpha, \beta\right) \propto \frac{\left(n_{-j}^{k,d_i^L} + \alpha\right)\left(n_{-j}^{w_{i,j}^L,k} + \beta\right)}{n_{-j}^k + V\beta}, \tag{6}$$

where $z_{i,j}^L$ represents the topic of the jth word $w_{i,j}^L$ in the ith text of L, $p(z_{i,j}^L = k \mid D_{-j}, \alpha, \beta)$ is the probability of word $w_{i,j}^L$ belonging to topic k conditioned on the whole set D after removing the current word $w_{i,j}^L$, n^{k,d_i^L} is the number of occurrences of topic k in text d_i^L, $n^{w_{i,j}^L,k}$ is the number of occurrences of word $w_{i,j}^L$ belonging to topic k in D, n^k is the number of occurrences of all words belonging to topic k in D, and V is the size of the vocabulary. Moreover, the subscript $._{-j}$ means topic k of word $w_{i,j}^L$ is removed from $\mathbf{z}_i^L$ where $\mathbf{z}_i^L$ represents the topics of all words in d_i^L. Here, K is the number of topics in D, and k=1,2,…,K. For instance, n_{-j}^{k,d_i^L} is obtained through removing the

(topic, word) combination at the jth word of the ith text of L. Unless noted otherwise, all counts are calculated based on the entire D. If all texts in D are long texts, HTTM is reduced to LDA.

For the ith short text, we can learn the latent variable z_i^S based on the following conditional distribution $p(z_i^S = k|D_{-i}, \beta)$,

$$p\left(z_i^S = k \mid D_{-i}, \beta\right) \propto \frac{\prod_{w \in d_i^S} \left(n_{-i}^{w,k} + \beta\right)}{\prod_{i=1}^{n_i^S} \left(n_{-i}^{k} + V\beta\right)}, \tag{7}$$

where the subscript $._{-i}$ means the ith short text of S is excluded from D, $n_{-i}^{w,k}$ is the number of occurrences of word w belonging to topic k in D without considering the ith short text, and n_{-i}^{k} is the number of occurrences of all words belonging to topic k in D excluding from d_i^S. Similarly, all counts are calculated based on the entire D. Here, HTTM is different from Unigrams [14] and DMM [9] about how to infer the latent variable z_i^S. z_i^S in Unigrams and DMM inclines to select a topic with more texts. Because our model need to evaluate the topics of long texts, we only consider the influence of words in short text for its topic.

Based on the fact that Dirichlet distribution is conjugate to multinomial distribution, we can get the posterior topic-word distribution of ϕ and text-topic distribution θ as follows:

$$\phi_k^w = \frac{n^{w,k} + \beta}{n^k + V\beta},$$

$$\theta_{d_i^L}^k = \frac{n^{k,d_i^L} + \alpha}{n_i^L + K\alpha},$$

$$\theta_{d_i^S}^k = \begin{cases} 1, & k = z_{d_i^S} \\ 0, & others \end{cases},$$

where ϕ_k^w is the probability of word w generated by topic k, and can be viewed as the distinction of word w to topic k, n_i^L is the number of words in long text d_i^L, and n^{k,d_i^L} is the number of words belonging to topic k in text d_i^L.

3.3 Reader-Aware Multi-Document Summarization

After obtaining the results (θ, ϕ) from news reports and user comments using HTTM, the topics cover not only the key topics of the news reports, but the concern of most readers. In this section, we discuss how to generate a summary using the topics discovered by HTTM, denoted as HTTM-based summarization. The basic idea of HTTM based summarization is that the sentences are chosen into the summary, which include more high probability words of important topics and are less similar to other sentences. The sentences of the summarization are only extracted from long texts, because short texts are very noisy, grammatically and informatively. This does not mean that the generated summary does not take the concern of most readers into consideration, because we calculate the scores of the sentences according to the topics mined from short texts and long texts. If the mined topics are mainly covered by short texts, only these sentences in long texts containing them have high scores.

Sentence score: Given all sentences s_m, $m \in \{1, M\}$, from L, a collection of long texts, and all hidden topics $\phi_k, k \in \{1, \ldots, K\}$. Assume that s_m belongs to long text d_i^L. Below, we will compute the probability of each sentences s_m of long text d_i^L given the topic ϕ_k, i.e., $\mathrm{P}(s_m \mid \phi_k)$.

Because sentences are often short text, we assume that each sentence only includes one topic, which is similar to short text. Thus, even words of the sentence s_m belonging to other topic are

forced to represent this topic:

$$p(s_m \mid \phi_k) = \prod_{w_p \in s_m} \phi_k^{w_p} \times \theta_{d_i^L}^k \times p\left(d_i^L\right), \tag{8}$$

where $p(s_m \mid \phi_k)$ is the probability of a sentence s_m given topic ϕ_k, $\phi_k^{w_p}$ denotes that the probability of word w_p is generated by topic ϕ_k, $\prod_{w_p \in s_m} \phi_k^{w_p}$ is the probability that the words of s_m belonging to topic k, $\theta_{d_i^L}^k$ that is the probability of topic k belonging to text d_i^L, and $p(d_i^L)$ is the probability of text d_i^L.

Since $\phi_k^{w_p} < 1$, Equation (8) uses the product of the probability of word w_p belonging to sentence s_m ($\prod_{w_p \in s_m} \phi_k^{w_p}$) that will penalize longer sentences. For example, two sentences s_1 and s_2 have q words in common. Let the length of s_1 be q and that of s_2 be $q + 1$. Thus, we have

$$\prod_{p=1}^{q} \phi_k^{w_p} > \prod_{p=1}^{q} \phi_k^{w_p} \times \phi_k^{w_{q+1}}$$

$$OR \quad p(s_1|k) > p(s_2 \mid k).$$

This will hold for all sentences and all topics, as a result the summary will consist of the shortest sentences in the documents. Because the probability distribution of topics for each text represent the weights of the topics and the topics can be regarded as a weighted mixture of words, we replace the product of probability of words using sum of the probability of words. By varying Equation (8), we get the following equation:

$$p(s_m \mid \phi_k) = \sum_{w_p \in s_m} \phi_k^{w_p} \times \theta_{d_i^L}^k \times p\left(d_i^L\right). \tag{9}$$

But, this equation brings a new problem. Longer sentences will have an advantage than shorter sentences, since longer sentences will always have higher probability measure since each word has a probability measure associated with it. However, due to the length limit of each summarization, the average weight of the longer sentence may not necessarily more than shorter sentence. Therefore, we normalize the probability value by dividing the length of the sentence n_{s_m}, we get the following equation:

$$p(s_m \mid \phi_k) = \frac{\sum_{w_p \in s_m} \phi_k^{w_p} \times \theta_{d_i^L}^k \times p\left(d_i^L\right)}{n_{s_m}}. \tag{10}$$

Thus, there is no guarantee that a longer sentence will be always preferred. The sentence will only be preferred if the added word in this sentence rises the recent total representation of the sentence toward the topic k.

To calculate $p(k)$ that represents the importance of this topic, we sum $\theta_{d_i^L}^k$ over long texts L,

$$p(k) = \sum_{l=1}^{L} \theta_{d_i^L}^k \times p\left(d_i^L\right), \tag{11}$$

where $\theta^k_{d_i^L}$ is text-topic distribution θ mined by HTTM. For text probabilities $p(d_i^L)$, one way of computing the probability is as given in LDA [7]:

$$P\left(d_i^L \mid \alpha, \beta\right) = \frac{\Gamma\left(\sum_{k=1}^K a_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int \left(\sum_{k=1}^K \left(\theta^k_{d_i^L}\right)^{\alpha_k - 1}\right) \left(\prod_{n=1}^{n_i^L} \sum_{k=1}^K \prod_{j=1}^V \left(\phi_k^{w_j} \theta^k_{d_i^L}\right)^{w_n^j}\right) d\theta_{d_i^L}. \tag{12}$$

Alternatively, if we know the probabilities of the texts, we can use that in our calculations instead of inferring them from the HTTM model. In our case, for the sake of simplicity, we have assumed that all documents are equiprobable. Thus, the probability of the text values do not make any difference during the calculation of the $p(s_m|k)$.

Hence, the score of the sentence s_m is signified as score(s_m) and is computed as follows:

$$score(s_m) = \sum_{k=1}^K p(s_m \mid k) \times p(k). \tag{13}$$

The score of each sentence is decided by the number of high probabilistic words of the core topics. We gain a conclusion that these sentences containing more high probabilistic words of core topics will have high scores.

Sentence selection: Considered both content coverage and non-redundancy, the aim is to generate a summary from multiple documents. In order to reduce redundancy in the summary, some approaches measure the similarity of next candidate sentence to that of previously selected ones, and select it if its similarity is below a threshold [34]. The most widespread strategy is maximal marginal relevance (MMR) [35]. At each iteration, MMR attempts to choose sentences that are unalike from the ones already selected.

The most widespread approach of calculating sentence relationship in MMR is the TFISF of the candidate terms [30, 35]. As sentences only have a few words comparing to text, most terms only have one occurrence in a sentence. Therefore, TF of TFISF cannot play an important role in calculating sentence relationship. So TFISF can scarcely make a distinction without considering the context information of each term. So we present a novel approach of calculating sentence relationship using topic distribution.

Each sentence has been represented using the topic distribution, namely $s_m = \{\theta^1_{s_m}, \theta^2_{s_m}, \ldots, \theta^K_{s_m}\}$, where each $\theta^k_{s_m}$ can be computed,

$$\theta^k_{s_m} = \frac{n^k_{s_m} + \alpha}{n_{s_m} + K \times \alpha},$$

where $n^k_{s_m}$ is the number of the words of sentence s_m belonging to topic k, and n_{s_m} is the number of words in sentence s_m.

We use an adjacency matrix R to describe the distance between two sentences. $R = [R_{ij}]_{MM}$ is defined as follows:

$$R_{ij} = D_{JS}(s_i, s_j) = \frac{1}{2} D_{KL}(s_i||s_j) + \frac{1}{2} D_{KL}(s_j||s_i), \tag{14}$$

where $D_{KL}(s_i||s_j)$ is the Kullback–Leibler divergence from s_i to s_j [36], and D_{JS} is Jensen–Shannon divergence [37]. Through this formula, the distance value between sentences is generated based on their topic distribution of the words in each sentence.

Let U be the set of sentences already chosen. The weight of each sentence is computed as

$$weight(s_i) = score(s_i) \times e^{\lambda \min_{s_j \in U} R_{ij}}, \tag{15}$$

where λ is a user-specified constant that adjusts the tradeoff between content coverage and non-redundancy: if $\lambda = 0$, the content coverage is highlighted; if $\lambda = 1$, only the non-redundancy is considered.

HTTM-based summarization adopts Equation (15) to order sentences. Through this function, we iteratively choose the top $|U|$ sentences with the highest scores, where the total length of these $|U|$ sentences must be below or equal to the length limitation. The pseudo-code of HTTM-based summarization is displayed in Algorithm 1.

ALGORITHM 1: HTTM-based summarization

Input: a set of news reports L consist of $\{s_1, s_2, \ldots, s_M\}$, a set of news comments S, topic number K, limitation length Lim;
Output: Summary (U)

```
{φ, θ} = HTTM(L, S, K)
for all sentences s_m ∈ L do
    for all topics k (k ∈ (1, ..., K)) do
        compute p(s_m|k) using Equation (10)
    end for
end for
for all topics k (k ∈ (1, ..., K)) do
    compute p(k) using Equation (11)
end for
for any two sentences s_i, s_j ∈ L do
    compute R_ij using Equation (14)
end for
U ← ∅
while len < Lim do
    highWeight ← 0 // recore the highest score
    id ← 0 // record the id of the sentence owns the highest score
    for each s_m ∈ L and m ∉ U do
        compute weight(s_m) using Equation (15)
        if weight(s_m) > highWeight then
            id = m, highWeight = weight(s_m)
        end if
    end for
    len = len + n_{s_id}
    U = U ∪ {s_id}
end while
return U
```

4 EXPERIMENTS

Since our HTTM-based summarization is based on the proposed topic modeling HTTM, we design experiments to answer the following questions:

Evaluation of HTTM: Does HTTM discover good latent topics than the existing topic modelings?

Evaluation of HTTM-based summarization: Does HTTM-based summarization system outperform state-of-the-art competitors?

For better evaluation of HTTM and HTTM-based summarization, we will adopt different datasets, metrics and baselines independently. We implement HTTM and HTTM-based summarization in JAVA.[1] All experiments were conducted on a Windows machine with an Intel 437 2.9GHz CPU and 8GB memory.

4.1 Datasets

In order to assess the effectiveness of topic modeling and summarization, we do experiments on the synthetic datasets and real-world datasets, respectively.

(1) Two synthetic datasets for topic modeling: *NIPS* [38] and *20 news group* [39]. The two datasets are composed of long texts. For better evaluation, similar to the previous papers [21, 40], we generate short texts by splitting part of documents in corpus into sentences, and fix the remaining documents, denoted as heterogeneous-length texts. Through changing long texts into heterogeneous-length texts, it is very useful for evaluating the quality of topic modeling. We will explain it in the section Evaluation Metrics.
(2) Two synthetic datasets for summarization: *DUC2002* and *DUC2004*. The standard benchmark DUC2002 and DUC2004 datasets are from Document Understanding Conference (DUC) for generic summarization evaluation. Because DUC2002 and DUC2004 cannot provide the corresponding comments, we also randomly choose a certain percentage of documents as long texts, and generate a set of short texts by splitting the remaining documents into sentences as comments. In this condition, long text is treated as news report and short text is used as reader comments. Here, the percentage of long texts of the total documents in DUC2002 and DUC2004 is set to 0.2.
(3) One real-world dataset for both topic modeling and summarization: *News&Tweet*. We choose one real-world dataset using in the paper [40], which includes eight hot events. They are "Tom Coughlin," "Oculus Rift," "SpaceX Rocket," "Donald Trump," "Windows 10," "Craig Strickland," "Bill Cosby," and "China stocks," respectively. In this experiment, we ask four human annotators to select several important sentences that can represent the core topics as human summarization. When extracting summaries, they take into account of the whole dataset including news reports and reader comments. The length for News&Tweet's summary is limited by 200 words.

For each corpus, we perform the following preprocessing. (1) We transfer all characters into lowercase; (2) we eradicate non-latin characters and stop words; (3) we eradicate words whose lengths are smaller than 3 or greater than 20. In addition, after preprocessing News&Tweet corpus, we only select these tweets whose length is greater than 5.

4.2 Evaluation Metrics

For better evaluation of topics and summarization, we choose different metrics.

Metrics of topic modeling: There are a lot of metrics to measure the coherence of topics in texts [41]. In this article, we choose the following two metrics.

(1) *Purity*: Purity is a metric presented by Quan et al. [21], which computes the coherence between gold-standard topics and discoverable topics. Since 20 News and NIPS datasets consist of long texts, LDA can discover good topics on the original datasets [42]. Therefore, we view the topics discovered from the original 20 News and NIPS datasets using LDA as gold-standard topics. We calculate purity by selecting the top T words for each topic,

[1] https://github.com/qiang2100/HTTM.git.

respectively, and match the top words with those from gold-standard topics:

$$Purity = \frac{1}{TK}\sum_{i} max_j |\ \Gamma_{z_i} \cap \Gamma_{g_j}\ |,$$

where z_i is a topic discovered from heterogeneous-length texts, g_j is a topic from gold-standard topics, and Γ_{z_i} and Γ_{g_j} are the sets of the top T words from topics z_i and g_j.

(2) *Qualitative and Quantitative Evaluation* [42]: Through choosing some exemplar topics learned by topic modelings on the News&Tweet dataset, we evaluate our model in a quantitative manner based on the coherence measure (CM) to assess how coherent the learned topics are. We choose the top 10 candidate words for each topic and request human annotators to assess whether they are relevant to the matching topic. To do this, annotators need to assess whether a topic is interpretable or not. If not, the 10 words of the topic are classified as irrelevant; else these words are recognized by annotators as relevant words for this topic. CM is defined as the ratio between the number of relevant words and the total number of candidate words. In our experiments, four graduate students joined in the recognized process.

Metric of summarization: We use the ROUGE toolkit [43] (version 1.5.5) to evaluate the summarization performance, which is adopted by DUC[2] as the official metric for document summarization. ROUGE evaluates the quality of a summary by counting the unit overlaps between the candidate summary and a set of human-generated summaries. The summary that achieves the highest ROUGE score is considered to be the most similar to the human-generated summary. As previously done in these literatures [1, 5], we choose the ROUGE-1 and ROUGE-2 metrics, and the *F*-measures of ROUGE-1 and ROUGE-2 are reported in this article. ROUGE-*N* measures the *N*-gram recall between a candidate summary and a set of reference summaries. ROUGE−*N* is computed as follows:

$$ROUGE-N = \frac{\sum_{S \in Sum_{ref}} \sum_{N-gram \in S} Count_{match}(N-gram)}{\sum_{S \in Sum_{ref}} \sum_{N-gram) \in S} Count(N-gram)} \tag{16}$$

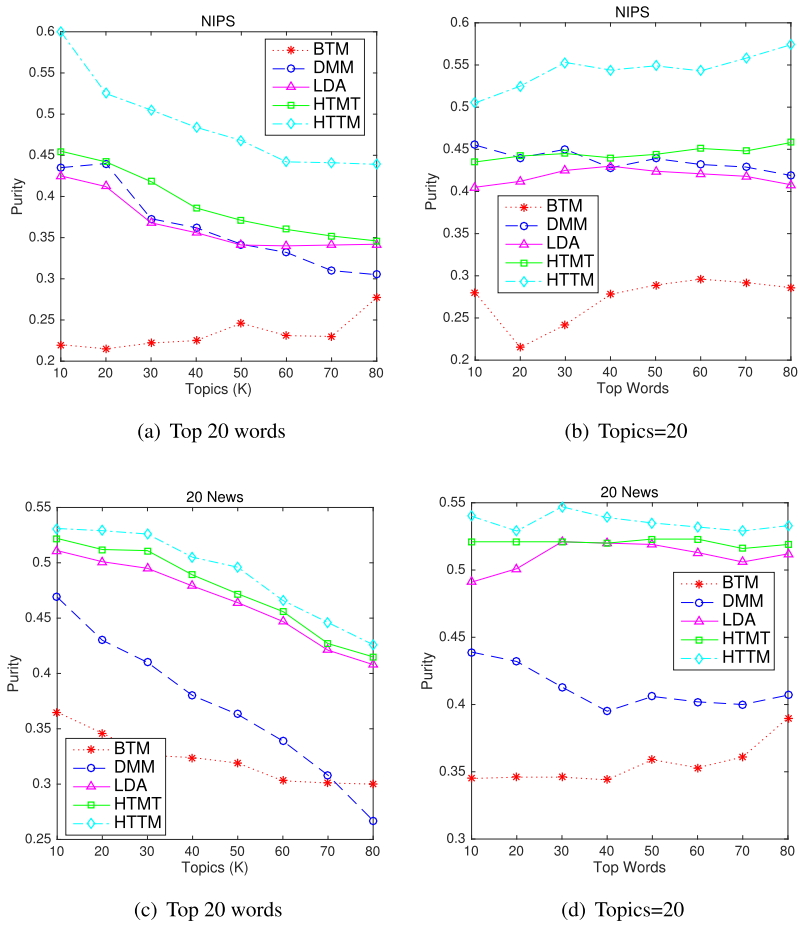
where N stands for the length of the N-gram, $Count_{match}(N\text{-gram})$ is the maximum number of N-grams co-occurring in candidate summary and the set of human-summaries, Count(N-gram) is the number of N-grams in the human summaries.

4.3 Comparison Methods

Methods of topic modeling: We compare our HDMM with other models including the following: (1) Long text topic model, LDA [8], which is the most widely used topic model. (2) Two state-of-the-art short text topic models, DMM [9] and BTM [11]. DMM uses the simple assumption that each text is generated from only one topic. BTM learns topics by directly modeling the generation of word co-occurrence patterns in the corpus. (3) One state-of-the-art heterogeneous-length topic model with Temporal Dynamics (HTMT) [3], which applies author topic model [44] on short texts and LDA on long texts.

Following [9], α and β of LDA and DMM are set as 0.1 and 0.1, respectively. In BTM, we use the recommended settings $\alpha = 50/K$, $\beta = 0.01$, and $\mu = 0.1$. Our model sets $\alpha = 0.1$, $\beta = 0.1$. The parameters of HTMT is set according to the original paper. The number of iterations is set to 2,000, which is generally sufficient for convergence. The percentage of long texts of the total documents in NIPS and 20 News is set to 0.2. In the experiments, we randomly generate 20 different heterogeneous-length datasets by fixing the percentage of long texts. We run each model 20 times

[2] http://duc.nist.gov.



(a) Top 20 words (b) Topics=20

(c) Top 20 words (d) Topics=20

Fig. 4. Experimental results on NIPS (left) and 20 news (right) data.

on all datasets, and report the mean. The length threshold is 20, namely if the length of a text is larger than 20, it is treated as a long text, else as a short text.

Methods of summarization: We compare our method with five summarization baselines. *Random* baseline randomly selects sentences for each topic. *LSA* [27], *NMF* [28], and *LDA* [6] based summarization are chosen to compare. The generated summary of the methods (Random, LSA, NMF, LDA) are shown from the news reports. *Sparse-code* based summarization [5] generates summaries using sparse code for RA-MDS, which considers the influence of reader comments.

4.4 Results of HTTM

Purity: By varying the number of topics K and top words when calculating purity, we verify the performance of topic modelings. Under fixing $T = 20$, topic numbers K from 10 to 80 are explored. Under fixing $K = 20$, top words from 10 to 80 are explored. The results are demonstrated in Figure 4.

From Figure 4, we find out that HTTM outperforms other methods. These results indicate that any one of the two assumptions (one text holds multiple topics or one topic) does not fit for heterogeneous-length texts. If we only adopt the simple assumption that one text holds only one topic, the model (DMM) will have poor performance on long texts. In contrast, if we only use the

complex assumption that one text holds multiple topics, the model (LDA) will not fit for short texts. Our model outperforms HTMT, although HTMT adopt complex assumption by applying author topic model for short texts and LDA for long texts.

Qualitative evaluation: Table 2 displays some paradigm topics inferred by the five methods on the News&Tweet corpus. The top 10 high probability words of each topic are envisaged, in which words that are deafening and lack of representativeness are dyed in bold. We see HTTM acquire more coherent topics with fewer deafening and worthless words than other models. The top 10 words of "Oculus Rift," and "Windows 10," learned by LDA are not related to the matching topic. Thus, LDA cannot work very well for short texts due to limited word co-occurrence patterns in short texts. DMM based on simple assumption that each text is sampled from one topic has the poorest results, since each long text sampled from one topic significantly decreases the performance. BTM suggests that two words co-occurring in a short text is assigned the identical topic cannot help increase the coherence of topic modeling as BTM overlooks the experience that all the words in a short text have a high chance from the identical topic. We can see that "Oculus Rift," learned by HTMT are not related to the matching topic, which indicates that our hypothesis for learning topics is more fit for heterogeneous-length texts. In summary, we can get that the topics discovered by HTTM are far better than those discovered by the baselines.

Quantitative evaluation: Table 3 displays the CM of topics discovered on News&Tweet corpus. We run each model 20 times on News&Tweet corpus, and report the mean. The coherence HTTM is 73%, which is significantly larger than BTM with 53.7% and DMM 34.7%. Compared with LDA and HTMT, HTTM still has a big enhancement. Therefore, HTTM yields better results compared to other models using quantitative evaluation.

Clustering: We further evaluate all models in clustering. To provide alternative metrics, one widely used metric, the normalized mutual information (NMI) is selected to evaluate the quality of a clustering solution [9, 16]. The value of NMI is always a number between 0 and 1, where 1 represents the greatest result and 0 means a random text partitioning. We train each model 20 times on each corpus and show the mean and standard deviation of their NMI values.

Table 4 displays the performance of all models on the News&Tweet corpus. First, we can get that HTTM outperforms all baselines including HTMT. It means that HTTM is the best algorithm for clustering heterogeneous-length texts. It indicates that our method is effective at clustering heterogeneous-length texts.

Influence of the percentage of long texts: When fixing $T = 20$ and $K = 20$ on NIPS corpus, we further explore the impact of the percentage of long texts on heterogeneous-length texts to the performance of all algorithms. Since more short texts (Twitter or Facebook) are generated online than long texts (News) in reality, the proportion of long texts in the whole texts is adjusted from 0.05 to 0.6 in the experiment. From Figure 5, we can get that HTTM outperforms other models. When growing the value of the percentage, the purity of HTTM, HTMT, LDA raises, and the purity of DMM and BTM drops. This reason is that more long texts are used to infer the topics as the percentage of long texts increases. But, when the proportion is lesser, LDA is poorer than DMM. This displays that LDA cannot word well for short texts.

4.5 Results of HTTM-Based Summarization

Results on DUC: As shown in Table 5, we can see that these baselines without considering reader comments are worse than the two methods (HTTM and sparse-coding) considering reader comments on DUC2002 dataset. In Table 6, all the ROUGE values of HTTM and sparse-code considering comments are also better than those ignoring comments with large gaps on DUC2004

Table 2. Topics Learned From News&Tweet Dataset

Topic	Method	Top words
Tom	LDA	coughlin giants coach season head **years** team york game super
	DMM	giants coughlin steps seasons **manage subscriptions alerts story artist charged**
	BTM	giants coughlin coach season head team **years** super game york
Coughlin	HTMT	coughlin giants coach season head **years** team york game super
	HTTM	coughlin giants coach season head team **years** york game super
Oculus	LDA	**windows oculus microsoft rift devices company headset january lumia hours**
	DMM	oculus rift **video** open headset reality free preorders **launch** virtual
	BTM	**windows microsoft oculus rift devices company news headset hours lumia**
Rift	HTMT	**windows oculus rift microsoft devices company headset open lumia hours**
	HTTM	oculus rift headset virtual reality january wednesday company touch price
SpaceX	LDA	rocket spacex falcon space landing launch musk stage **company** satellites
	DMM	**windows cosby year trump microsoft rocket china monday spacex time**
	BTM	rocket spacex falcon space launch landing musk stage **company** satellites
Rocket	HTMT	rocket spacex falcon space landing launch musk stage **company** satellites
	HTTM	rocket spacex falcon space landing launch musk stage **company** satellites
Donald	LDA	trump campaign donald iowa cruz republican clinton state hampshire states
	DMM	**coughlin bill cosby giants trump donald odell beckham coach video**
	BTM	trump campaign iowa cruz republican donald hampshire clinton rubio states
Trump	HTMT	trump campaign donald iowa cruz republican clinton state hampshire states
	HTTM	trump campaign donald iowa cruz republican clinton state hampshire states
Windows	LDA	**email comments post news today comment account facebook access badge**
	DMM	windows microsoft **keys** mobile **encryption** tablet users **acer liquid jade**
	BTM	**died year attack saudi january oregon fire iran star best**
10	HTMT	windows microsoft mobile devices tablet lumia company percent running december
	HTTM	windows microsoft devices lumia company mobile percent users running december
Craig	LDA	strickland craig body oklahoma singer missing morland country lake helen
	DMM	strickland craig singer missing country hope **great** dead church **friend**
	BTM	strickland craig body oklahoma morland missing singer lake helen country
Strickland	HTMT	strickland craig body oklahoma singer missing morland lake helen country
	HTTM	strickland craig body oklahoma singer missing morland country lake monday
Bill	LDA	cosby bill women assault case camille sexual court constand charged
	DMM	**giants coughlin steps seasons manage subscriptions alerts story artist charged**
	BTM	cosby bill women case assault camille sexual court constand **years**
Cosby	HTMT	cosby bill women assault case camille sexual court constand charged
	HTTM	cosby bill women assault case camille sexual court **years** constand
China	LDA	china stocks trading markets percent **year** market stock chinese investors
	DMM	china stocks hong kong policy data optimism open inflation **morning**
	BTM	china percent trading markets **year** stocks market stock chinese investors
stocks	HTMT	china stocks trading markets percent **year** market stock **morning** investors
	HTTM	china stocks trading markets percent **year** market stock chinese investors

Table 3. CM (%) on News&Tweet Dataset

Method	Annotator1	Annotator2	Annotator3	Annotator4	Mean	Standard deviation
LDA	70	61	55	57	60.7	6.6
DMM	34	30	35	40	34.7	4.1
BTM	56	47	56	56	53.7	4.5
HTMT	69	66	61	64	65	3.4
HTTM	**72**	**73**	**74**	**73**	**73**	**0.8**

Bold face highlights the best number.

Table 4. NMI Values on News&Tweet Dataset

Method	NMI
LDA	0.8350 ± 0.0349
DMM	0.5194 ± 0.0364
BTM	0.8786 ± 0.0281
HTMT	0.8972 ± 0.0231
HTTM	**0.9235 ± 0.0303**

Bold face highlights the best number.

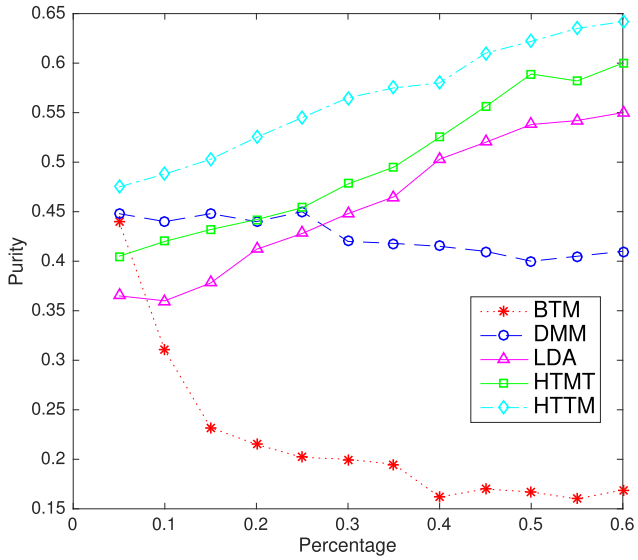


Fig. 5. Comparison of the models with different percentages of long texts for NIPS dataset.

dataset. It means that read comments are useful for MDS. The performance of HTTM is better than other methods in terms of the results of ROUGE-1 and ROUGE-2 metrics, except the ROUGE-1 on DUC2002. On DUC2002, HTTM achieves the best ROUGE-2 (0.2015) and the second-best ROUGE-1 (0.4765). On DUC2004, HTTM achieves the best ROUGE-1 (0.3867) and ROUGE-2 (0.0856) value. The closest method to HTTM is sparse-coding that also takes advantage of comments to calculate the salience of the text units when choosing sentences. It illustrates that our method can generate good summary, because the topics mined by HTTM capture the events being covered by news reports and reader comments.

Table 5. ROUGE Values for Methods Obtained with DUC2002 Dataset

Method	ROUGE-1	ROUGE-2
Random	0.3878	0.1196
LSA	0.3952	0.1257
NMF	0.4023	0.1289
LDA	0.4019	0.1292
Sparse-code	**0.4832**	0.1989
HTTM	0.4765	**0.2015**

Table 6. ROUGE Values for Methods Obtained with DUC2004 Dataset

Method	ROUGE-1	ROUGE-2
Random	0.3227	0.0639
LSA	0.3273	0.0648
NMF	0.3324	0.0681
LDA	0.3370	0.0696
Sparse-code	0.3832	0.0815
HTTM	**0.3867**	**0.0856**

Table 7. ROUGE Values for Methods Obtained with News&Tweet Dataset

Method	ROUGE-1	ROUGE-2
Random	0.3428	0.0662
LSA	0.3694	0.0750
NMF	0.3827	0.0782
LDA	0.3761	0.0776
Sparse-coding	0.4063	0.0930
HTTM	**0.4164**	**0.107**

Results on News&Tweet: Table 6 shows the results of HTTM compared with other automatic methods in terms of ROUGE-1 and ROUGE-2 on News&Tweet dataset. Similar to the results on DUC2002 and DUC2004, we can see that the two methods (HTTM and Sparse-coding) considering comments achieve the best performance. The main reason is that these baselines only consider the knowledge of news reports and overlook the content of reader comments. HTTM performs a little better than Sparse-coding, because sparse-coding gives no considerations for the properties of heterogeneous-length corpus. In addition, the topics mined by HTTM can provide an intuitive way to express the core content of the whole collection. Through the results of Table 7, the effectiveness and feasibility of the improved algorithm is verified again.

Influence of the reader comments for HTTM method: In our above experiments, we extract topics from both long texts and short texts. In this experiment, we conduct some experiments by considering comments and ignoring comments for verifying the importance of reader comments for summarization. Therefore, we generate a new dataset "News," by ignoring short texts (Tweet) from

Table 8. ROUGE Values for HTTM

Dataset	ROUGE-1	ROUGE-2
DUC2002-noComments	0.4072	0.1315
DUC2002	0.4765	0.2015
DUC2004-noComments	0.3386	0.0714
DUC2004	0.3867	0.0856
News	0.3784	0.0768
News&Tweet	0.4164	0.107

News&Tweet. By ignoring short texts from two synthetic datasets DUC2002 and DUC2004, we generated two revised datasets that are denoted as DUC2002-nocomments and DUC2004-noComment. The results by ROUGE evaluation are given in Table 8. All the ROUGE values of HTTM considering comments significantly perform better than those ignoring comments. It expresses that considering comments for our model can improve the performance of the summarization, which verifies the importance of short texts for heterogeneous-length corpus.

5 CONCLUSION

In this article, we use latent topics mined by topic modelings to extract a summary from news reports through considering reader comments. Our key theme uses reader comments as an additional source to influence the summary from news reports for capturing the demand of the readers, and further customizes the summary extraction process to the users so the extracted summaries are specific to the underlying aspects of each event. In this article, we need to deal with heterogeneous-length texts including news reports and reader comments, where news reports are often long, and reader comments are often short. To improve the efficiency of topic modeling, we first proposed an efficient topic modeling algorithm, HTTM, to discover latent topics from heterogeneous-length texts. The key idea of HTTM adopts the two different assumptions simultaneously to infer the topics for short texts and long texts, respectively. We further calculate the weights of the sentences, which can capture more high probability words of important topics and are less similar to other sentences for summary extraction. Experiments and comparisons demonstrate that HTTM is effective in improving the quality of extracted topics, and the proposed multiple-document summarization algorithm outperforms existing multiple-document summarization algorithms.

REFERENCES

[1] Rasim M. Alguliev, Ramiz M. Aliguliyev, and Nijat R. Isazade. 2013. Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications* 40, 5 (2013), 1675–1689.

[2] Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah. 2013. Multi-document summarization based on the Yago ontology. *Expert Systems with Applications* 40, 17 (2013), 6976–6984.

[3] Long Chen, Huaizhi Zhang, Joemon M. Jose, Haitao Yu, Yashar Moshfeghi, and Peter Triantafillou. 2017. Topic detection and tracking on heterogeneous information. *Journal of Intelligent Information Systems* 10–17 (2017), 1–23.

[4] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. A probabilistic model for bursty topic discovery in microblogs. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*. 353–359.

[5] Piji Li, Lidong Bing, Wai Lam, Hang Li, and Yi Liao. 2015. Reader-aware multi-document summarization via sparse coding. In *Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI'15)*. 1270–1276.

[6] Rachit Arora and Balaraman Ravindran. 2008. Latent Dirichlet allocation based multi-document summarization. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data*. ACM, 91–97.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3 (2003), 993–1022.

[8] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl. 1 (2004), 5228–5235.

[9] Jianhua Yin and Jianyong Wang. 2014. A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 233–242.

[10] Rumi Ghosh and Asur Sitaram. 2013. Mining information from heterogenous sources: A topic modelling approach. In *Proc. of the MDS Workshop at the 19th ACM SIGKDD (MDS-SIGKDD'13)*.

[11] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26, 12 (2014), 2928–2941.

[12] Xin Wang, Ying Wang, Wanli Zuo, and Guoyong Cai. 2015. Exploring social context for topic identification in short and noisy texts. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

[13] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 50–57.

[14] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39, 2–3 (2000), 103–134.

[15] Guan Yu, Ruizhang Huang, and Zhaojun Wang. 2010. Document clustering via Dirichlet process mixture model with feature selection. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. ACM, 763–772.

[16] Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi. 2013. Dirichlet process mixture model for document clustering with feature partition. *IEEE Transactions on Knowledge and Data Engineering* 25, 8 (2013), 1748–1759.

[17] Jianhua Yin and Jianyong Wang. 2016. A text clustering algorithm using an online clustering scheme for initialization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. 1995–2004.

[18] Jipeng Qiang, Yun Li, Yunhao Yuan, and Xindong Wu. 2018. Short text clustering based on Pitman–Yor process mixture model. *Applied Intelligence* 48, 7 (2018), 1802–1812.

[19] Yuan Zuo, Jichang Zhao, and Ke Xu. 2016. Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems* 48, 2 (2016), 379–398.

[20] Wu Wang, Houquan Zhou, Kun He, and John E. Hopcroft. 2017. Learning latent topics from the word co-occurrence network. In *Proceedings of National Conference of Theoretical Computer Science*. 18–30.

[21] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th International Conference on Artificial Intelligence*. 2270–2276.

[22] Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2105–2114.

[23] Yang Yang, Feifei Wang, Junni Zhang, Jin Xu, and S. Yu Philip. 2018. A topic model for co-occurring normal documents and short texts. *World Wide Web* 21, 2 (2018), 487–513.

[24] Ruifang He, Jiliang Tang, Pinghua Gong, Qinghua Hu, and Wang Bo. 2016. Multi-document summarization via group sparse learning. *Information Sciences An International Journal* 349, C (2016), 12–24.

[25] Yu-Tong Wu, Xue-Feng Li, Yue Xu, and Wei Wang. 2016. Mining topically coherent patterns for unsupervised extractive multi-document summarization. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. 129–136.

[26] Jin Ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Compressive document summarization via sparse optimization. In *Proceedings of the International Conference on Artificial Intelligence*.

[27] Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 19–25.

[28] Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. 2009. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management* 45, 1 (2009), 20–34.

[29] Ji-Peng Qiang, Ping Chen, Wei Ding, Fei Xie, and Xindong Wu. 2016. Multi-document summarization using closed patterns. *Knowledge-Based Systems* 99 (2016), 28–38.

[30] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management* 40, 6 (2004), 919–938.

[31] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.

[32] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 255–264.

[33] Leonhard Hennig, Winfried Umbrath, and Robert Wetzker. 2008. An ontology-based approach to text summarization. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 3 (WI-IAT'08)*. IEEE, 291–294.

[34] K. Sarkar. 2010. Syntactic trimming of extracted sentences for improving extractive multi-document summarization. *Journal of Computing* 2 (2010), 177–184.
[35] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–336.
[36] Solomon Kullback. 1997. *Information Theory and Statistics*. Courier Corporation.
[37] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Vol. 999. MIT Press.
[38] Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. 2004. Euclidean embedding of co-occurrence data. In *Proceedings of Advances in Neural Information Processing Systems*. 497–504.
[39] Thorsten Joachims. 1996. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. Technical Report. DTIC Document.
[40] Jipeng Qiang, Ping Chen, Wei Ding, Tong Wang, Fei Xie, and Xindong Wu. 2016. Topic discovery from heterogeneous texts. In *Proceedings of the IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI'16)*. IEEE, 196–203.
[41] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 262–272.
[42] Pengtao Xie, Diyi Yang, and Eric P. Xing. 2015. Incorporating word correlation knowledge into topic modeling. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics*.
[43] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*, Vol. 8. Barcelona, Spain.
[44] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 306–315.

Received July 2018; revised May 2019; accepted May 2019