HIERARCHICAL THEME AND TOPIC MODEL FOR SUMMARIZATION

Jen-Tzung Chien and Ying-Lan Chang

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30010, ROC
jtchien@nctu.edu.tw

ABSTRACT

This paper presents a hierarchical summarization model to extract representative sentences from a set of documents. In this study, we select the thematic sentences and identify the topical words based on a hierarchical theme and topic model (H2TM). The latent themes and topics are inferred from document collection. A tree stick-breaking process is proposed to draw the theme proportions for representation of sentences. The structural learning is performed without fixing the number of themes and topics. This H2TM is delicate and flexible to represent words and sentences from heterogeneous documents. Thematic sentences are effectively extracted for document summarization. In the experiments, the proposed H2TM outperforms the other methods in terms of precision, recall and F-measure.

Index Terms— Topic model, structural learning, Bayesian nonparametrics, document summarization

1. INTRODUCTION

As the internet grows prosperously, the online web documents have been too redundant to browse and search efficiently. Automatic summarization becomes crucial for browsers to capture themes and concepts in a short time. Basically, there are two kinds of solutions to summarization. *Abstraction* is to rewrite summary for a document while *extraction* is to extract the representative sentences for a summary. Abstraction is usually difficult and arduous, so mostly we focus on extraction. However, a good summary system should reflect diverse topics of documents and keep redundancy to a minimum. Extraction likely leads to a summary with too coherent topics.

In the literature, the unsupervised learning via *probabilistic topic model* [1] has been popular for document categorization [2], speech recognition [3], text segmentation [4], and image analysis [1]. The latent semantic topics are learnt from a bag of words. Such topic model can capture the salient themes embedded in data collection and work for document summarization [5]. However, topic model based on latent Dirichlet allocation (LDA) [2] was constructed as a finite-dimensional mixture representation which assumed that 1) number of topics was fixed, and 2) topics were independent. The hierarchical Dirichlet process (HDP) [6] and the nested Chinese restaurant process (nCRP) [7][8] were proposed to conduct structural learning to relax these two assumptions.

HDP [6] is a Bayesian nonparametric extension of LDA where the representation of documents is allowed to grow structurally as more data are observed. Each word token within a document is drawn from a mixture model where the hidden topics are shared across documents. Dirichlet process (DP) is realized to find flexible data partitions or provide the nonparametric prior over number of topics for each document. The base measure for the child Dirichlet processes (DPs) is itself drawn from a parent DP. On the other hand, nCRP [7][8] explores the topic hierarchies with flexible extension of infinite branches and infinite layers. Each document selects a tree path with nodes containing topics in different sharing conditions. All words in the document are represented by using these topics.

In this study, we develop a hierarchical tree model for representation of sentences from heterogeneous documents. Using this model, each path from root node to leaf node covers from general theme to individual theme. These themes contain coherent information but in varying degrees of sharing. The brother nodes expand the diversity of themes from different sentences. This model does not only group sentences into a node in terms of its theme, but also distinguish their concepts by means of different levels. A structural stick-breaking process is proposed to draw a subtree path and determine a variety of theme proportions. We conduct the task of multi-document summarization where the sentences are selected across documents with a diversity of themes and concepts. The number of latent components and the dependency between these components are flexibly learnt from the collected data. Further, the words of the sentences inside a node are represented by a topic model which is drawn by DP. All the topics from different nodes are shared under a global DP. We propose Bayesian nonparametric approach to structural learning of latent topics and themes from the observed words and sentences, respectively. This approach is applied for concept-based summarization over multiple text documents.

978-1-4799-1180-6/13/$31.00 ©2013 IEEE

2. BAYESIAN NONPARAMETRIC LEARNING

2.1. HDP and nCRP

There have been many Bayesian nonparametric approaches developed for discovering a countably infinite number of latent features in a variety of real-world data. Bayesian inference is performed by integrating out the infinitely many parameters. HDP [6] conducts Bayesian nonparametric representation of documents or grouped data where each document or group is associated with a mixture model. Words in different documents share a global mixture model. Using HDP, each document d is associated with a draw from a DP G_d, which determines how much each member of a shared set of mixture components contributes to that document. The base measure of G_d is itself drawn from a global DP G_0 which ensures that there is a set of mixtures shared across data. Each distribution G_d governs the generation of words for a document d. The strength parameter α_0 determines the proportion of a mixture in a document d. The document distribution G_d is generated by $G_0 \sim \text{DP}(\gamma, H)$ and $G_d \sim \text{DP}(\alpha_0, G_0)$ where $\{\gamma, \alpha_0\}$ and H denote the strength parameters and the base measure, respectively. HDP is developed to represent a bag of words from a set of documents through nonparametric prior G_0. In [7][8], the nCRP was proposed to conduct Bayesian nonparametric inference of topic hierarchies and learn the deeply branching trees from data collection. Using this hierarchical LDA (hLDA), each document was modeled by a path of topics along a random tree where the hierarchically-correlated topics from global topics to specific topics were extracted. In general, HDP and nCRP could be implemented by using stick-breaking process and Chinese restaurant process. The approximate inference algorithms via Markov chain Monte Carlo (MCMC) [6][7][8] and variational Bayesian [9][10] were developed.

2.2. Stick-Breaking Process

Stick-breaking process is designed to implement infinite mixture model according to a DP. Beta distribution is introduced to draw binary variables for stick-breaking into left segment and right segment. A random probability measure G is first drawn from a DP with base measure H using a sequence of beta variates. Using this process, a stick of unit length is partitioned at a random location. The left segment is denoted by θ_1. The right segment is further partitioned at a new location. The partitioned left segment is denoted by θ_2. We continue this process by generating the left segment θ_i and breaking the right segment at each step i. Stick-breaking depends on a random value drawn from H which is seen as center of probability measure. The distribution over sequence of proportions $\{\theta_1, \cdots, \theta_i\}$ is called GEM distribution which provides a distribution over infinite partitions of unit interval [11]. In [12], a tree stick-breaking process was proposed to infer a tree structure. This method interleaved two stick-breaking procedures. The first has beta variates for depth which determine the size of a given node's partition as a function of depth. The second has beta variates for branch which determine the branching probabilities. Interleaving two procedures could partition the unit interval into a tree structure.

3. HIERARCHICAL THEME AND TOPIC MODEL

In this study, a hierarchical theme and topic model (H2TM) is proposed for representation of sentences and words from a collection of documents based on Bayesian nonparametrics.

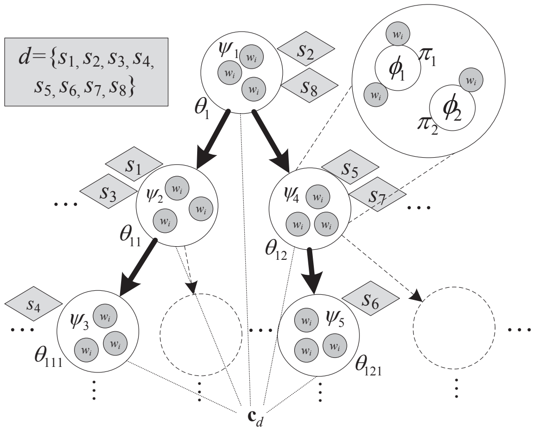


Fig. 1. A tree structure for representation of words, sentences and documents. Thick arrows denote the tree paths $\mathbf{c}_d$ drawn for eight sentences of a document d. Dark rectangle, diamonds and circles denote the observed document, sentences and words, respectively. Each sentence s_j is assigned with a theme variable ψ_l at a tree node along tree paths with probability θ_{dl} while each word w_i in tree node is assigned with a topic variable ϕ_k with probability π_{lk}.

3.1. Model Description

H2TM is constructed by considering the structure of a document where each document consists of a "bag of sentences" and each sentence consists of a "bag of words". Different from the infinite topic model using HDP [6] and the hierarchical topic model using nCRP [7][8], we propose a new tree model for representation of a "bag of sentences" where each sentence has *variable length of words*. A two-stage procedure is developed for document representation as illustrated in Figure 1. In the first stage, each sentence s_j of a document is drawn from a *mixture of theme model* where the themes are shared for all sentences from a collection of documents. The theme model of a document d is composed of the themes along its corresponding tree paths $\mathbf{c}_d$. With a tree structure of themes, the unsupervised grouping of sentences into different layers is constructed. In the second stage, each word w_i of the sentences allocated in a tree node is drawn by an individual *mixture of topic model*. All topics from different nodes are drawn using a global topic model.

Using H2TM, we assume that the words of the sentences in a tree node given topic k are conditionally independent and drawn from a topic model with infinite topics $\{\phi_k\}_{k=1}^{\infty}$. The sentences in a document given theme l are conditionally independent and drawn from a theme model with infinite themes $\{\psi_l\}_{l=1}^{\infty}$. The document-dependent theme proportions $\{\theta_{dl}\}_{l=1}^{\infty}$ and the theme-dependent topic proportions $\{\pi_{lk}\}_{k=1}^{\infty}$ are introduced. Given these proportions, each word w_i is drawn from a mixture model of topics $\sum_k \pi_{lk} \cdot \phi_k$ while each sentence s_j is sampled from a mixture model of themes $\sum_l \theta_{dl} \cdot \psi_l$. *Since a theme for sentences is represented by a mixture model of topics for words, we accordingly bridge the relation between themes and topics via* $\psi_l \sim \sum_k \pi_{lk} \cdot \phi_k$.

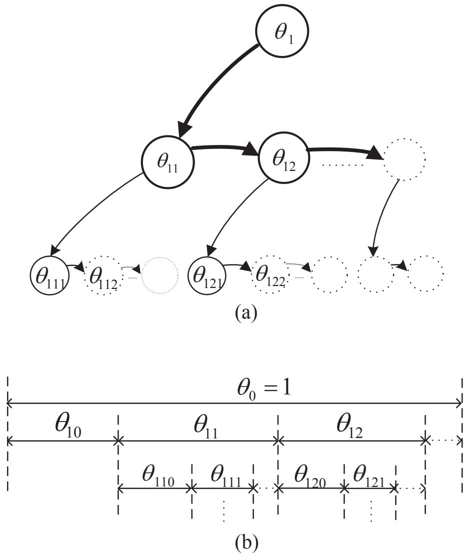
3.2. Sentence-Based nCRP

In this study, a sentence-based tree model with infinite nodes and branches are estimated to conduct unsupervised structural learning and select semantically-rich sentences for document summarization. A sentence-based nCRP (snCRP) is proposed to construct a tree model where root node contains general theme and leaf node conveys a specific theme for sentences. Different from previous word-based nCRP [7][8] where topics along a single tree path are selected to represent all words of a document, the snCRP is exploited to represent all sentences of a document based on the themes which are from multiple tree paths or equivalently from a *subtree* path. It is because that the variation of themes does exist in heterogeneous documents. The conventional word-based nCRP using GEM distribution should be extended to the snCRP using tree-based GEM (*tree*GEM) distribution by considering multiple paths for document representation. A tree stick-breaking process is proposed to draw a subtree path and determine the theme proportions for representation of all sentences in a document.

A new scenario is described as follows. There are infinite number of Chinese restaurants in a city. Each restaurant has infinite tables. A tourist visits the first (root) restaurant where each of its tables has a card showing the next restaurant which is arranged in the second layer of this tree. Such visit repeats infinitely. Each restaurant is associated with a tree layer, and each table has its unique label. The restaurants in a city are organized into an infinitely-branched and infinitely-deep tree structure. Model construction for H2TM is summarized by

1. For each theme l
   (a) Draw a topic model $\phi_k \sim G_0$.
   (b) Draw topic proportions $\pi_l | \{\alpha_0, \lambda_0\} \sim \text{DP}(\alpha_0, \lambda_0)$.
   (c) Theme model is generated by $\psi_l \sim \sum_k \pi_{lk} \cdot \phi_k$.
2. For each document $d \in \{1, \cdots, D\}$
   (a) Draw a subtree path $\mathbf{c}_d = \{c_{dj}\} \sim \text{snCRP}(\gamma_s)$.
   (b) Draw theme proportions over path $\mathbf{c}_d$ by tree stick-breaking $\theta_d | \{\alpha_s, \lambda_s\} \sim tree\text{GEM}(\alpha_s, \lambda_s)$.
   (c) For each sentence s_j
      i. Choose a theme label $z_{sj} = l | \theta_d \sim \text{Mult}(\theta_d)$.
      ii. For each word w_i
         A. Choose a topic label based on topic proportion of theme l, i.e. $z_{wi} = k | \pi_l \sim \text{Mult}(\pi_l)$.
         B. Draw a word based on topic z_{wi} by $w_i | \{z_{wi}, \phi_k\} \sim \text{Mult}(\phi_{z_{wi}})$.

The hierarchical grouping of sentences is accordingly obtained through a nonparametric tree model based on snCRP. Each tree node stands for a theme. A sentence s_j is determined by a theme model ψ_l. In what follows, we address how proportions θ_d of theme $z_{sj} = l$ are drawn for representation of words w_i of sentences s_j in document d.



(a)

(b)

Fig. 2. Illustrations for (a) tree stick-breaking process, and (b) hierarchical theme proportions.

3.3. Tree Stick-Breaking Process

Traditional GEM is used to provide a distribution over infinite proportions. Such distribution is not suitable for characterizing a tree structure with dependencies between parent nodes and child nodes. To cope with this issue, we present a new snCRP by conducting a tree stick-breaking (TSB) process where the theme proportions along with a subtree path are drawn. The subtree path is chosen to reveal a variety of subjects from different sentences while different levels are built to characterize the hierarchy of aspects. Each sentence is assigned by a node with theme proportion determined by all nodes in the selected subtree path $\mathbf{c}_d$.

Interestingly, the proposed TSB process is a special realization of the tree-structured stick-breaking process given in [12]. This process is specialized to draw theme proportions $\theta_d = \{\theta_{dl}\}$ for a document d subject to $\sum_{l=1}^{\infty} \theta_{dl} = 1$ based on a tree model with infinite nodes. TSB process is described as follows. We consider a set of a parent node and its child nodes that are connected as shown by thick arrows in Figure 2(a). Let l_a denote an ancestor node and $l_c = \{l_{a1}, l_{a2}, \cdots\}$ denote its child nodes. TSB is run for each set of nodes $\{l_a, l_c\}$ in a recursive fashion. Figures 2(a) and 2(b) illustrate how the tree structure in Figure 1 is constructed. Figure 2(b) shows how theme proportions are inferred by TSB. The theme proportion $\theta_{l_a 0}$ in the beginning child node denotes the initial fragment of node l_a when proceeding stick-breaking process for its child nodes l_c. Here, $\theta_0 = 1$ denotes the initial unit length, $\theta_1 = \nu_1$ denotes the first fragment of stick for root node and $1 - \nu_1$ denotes the remaining fragment of the stick. Given the *tree*GEM parameters $\{\alpha_s, \lambda_s\}$, the beta variable $\nu_u \sim \text{Beta}\left(\alpha_s\lambda_s, \alpha_s(1-\lambda_s)\right)$ of a child node $l_u \in \Omega_{l_c}$ is first drawn. The probability of generating this draw is calculated by $\nu_u \prod_{v=0}^{u-1} (1 - \nu_v)$. This probability is then multiplied by the theme proportion θ_{l_a} of ancestor node l_a so as to find theme proportion for its child nodes $l_u \in \Omega_{l_c}$. We can recursively calculate the theme proportion by

$$\theta_{l_u} = \theta_{l_a} \nu_u \prod_{v=0}^{u-1} (1 - \nu_v), \quad \text{for } l_u \in \Omega_{l_c}. \tag{1}$$

Therefore, a tree model is constructed without limitation of tree layers and branches. We improve the efficiency of tree stick-breaking method [12] by adopting a single set of beta parameters $\{\alpha_s, \lambda_s\}$ for stick-breaking towards depth as well as branch. Using this process, we draw a global theme proportions for sentences in different documents d by using scaling parameter γ_s and then determine a subtree path for all sentences s_j in document d via snCRP by $\mathbf{c}_d = \{c_{dj}\} \sim \text{snCRP}(\gamma_s)$. A tree stick-breaking process is performed to sample the theme proportions $\theta_d \sim tree\text{GEM}(\alpha_s, \lambda_s)$.

3.4. HDP for Words

After having the hierarchical grouping of sentences based on snCRP, we treat the words corresponding to a node of theme l as grouped data and conduct HDP using the grouped data from different tree nodes. The topic model is then constructed and utilized to draw individual words. Importantly, each theme is represented by a mixture model of topics $\psi_l \sim \sum_k \pi_{lk} \cdot \phi_k$. HDP is applied to infer word distributions and topic proportions. The standard stick-breaking process is applied to infer topic proportions for DP mixture model based on GEM distribution. The words of a tree node corresponding to theme l is generated by

$$\begin{gathered} \lambda_0 \sim \text{GEM}(\gamma_w), \quad \pi_l \sim \text{DP}(\alpha_0, \lambda_0), \quad \phi_k \sim G_0 \\ z_{wi}|\pi_l \sim \text{Mult}(\pi_l), \quad w_i|\{z_{wi}, \phi_k\} \sim \text{Mult}(\phi_{z_{wi}}) \end{gathered} \tag{2}$$

where λ_0 is a global prior for tree nodes, π_l is the topic proportion for theme l, ϕ_k is the kth topic, α_0 and γ_w are the strength parameters for DP. At last, the *snCRP compound HDP* is fulfilled to establish the hierarchical theme and topic model (H2TM).

4. MODEL INFERENCE

The approximate inference using Gibbs sampling is developed to infer posterior parameters or latent variables for H2TM. Each latent variable is iteratively sampled by a posterior probability with the condition on the observations and all the other latent variables. We sample tree paths $\mathbf{c}_d = \{c_{dj}\}$ for different sentences of document d. Each sentence s_j is grouped into a tree node with theme l which is sampled by proportions θ_d under a subtree path $\mathbf{c}_d$ through snCRP. Each word w_i of a sentence is assigned by the topic k which is sampled via HDP.

4.1. Sampling of Tree Paths

A document is treated as "a bag of sentences" for path sampling in proposed snCRF. To do so, we iteratively sample tree paths $\mathbf{c}_d$ for words $\mathbf{w}_d$ in document d consisting of sentences $\{\mathbf{w}_{dj}\}$. Sampling tree paths is performed according to the posterior probability

$$\begin{aligned} p(c_{dj}|\mathbf{c}_{d(-j)}, \mathbf{w}_d, z_{sj}, \psi_l, \gamma_s) \propto p(c_{dj}|\mathbf{c}_{d(-j)}, \gamma_s) \\ \times p(\mathbf{w}_{dj}|\mathbf{w}_{d(-j)}, z_{sj}, \mathbf{c}_d, \psi_l) \end{aligned} \tag{3}$$

where $\mathbf{c}_{d(-j)}$ denotes the paths of all sentences in document d except sentence s_j. The notation "-" denotes the self-exception. In (3), γ_s is Dirichlet prior parameter for global theme proportions. The first term in right-hand-side (RHS) calculates the probability of choosing a path for a sentence. This probability is determined by applying CRP [8] where the jth sentence chooses either an occupied path h by $p(c_{dj} = h|\mathbf{c}_{d(-j)}, \gamma_s) = \frac{f_{d(c_{dj}=h)}}{f_{d\cdot} - 1 + \gamma_s}$ or or a new path by $p(c_{dj} = \text{new}|\mathbf{c}_{d(-j)}, \gamma_s) = \frac{\gamma_s}{f_{d\cdot} - 1 + \gamma_s}$ where $f_{d(c_{dj}=h)}$ denotes the number of sentences in document d that are allocated along tree path h. Path h is selected for sentence $\mathbf{w}_{dj}$. The second term in RHS of (3) can be calculated by referring [7][8].

4.2. Sampling of Themes

Given the current path c_{dj} selected via snCRP by using words $\mathbf{w}_{dj}$, we sample a tree node at level ℓ or equivalently sample a theme l according to the posterior probability given current values of all other variables

$$\begin{aligned} p(z_{sj} = l|\mathbf{w}_d, \mathbf{z}_{s(-j)}, c_{dj}, \alpha_s, \lambda_s, \psi_l) \propto \\ p(z_{sj} = l|\mathbf{z}_{s(-j)}, c_{dj}, \alpha_s, \lambda_s) p(\mathbf{w}_{dj}|\mathbf{w}_{d(-j)}, \mathbf{z}_s, \psi_l) \end{aligned} \tag{4}$$

where $\mathbf{z}_s = \{z_{sj}, \mathbf{z}_{s(-j)}\}$. The number of theme is unlimited. The first term in RHS of (4) is a distribution over levels derived as an expectation of *tree*GEM which is implemented via TSB process and is calculated via a product of beta variables $\nu_u \sim \text{Beta}\left(\alpha_s\lambda_s, \alpha_s(1-\lambda_s)\right)$ along path c_{dj}. The second term calculates the probability of sentence $\mathbf{w}_{dj}$ given the theme model ψ_l.

4.3. Sampling of Topics

According to HDP, we apply stick-breaking construction to draw topics for words in different tree nodes. We view words $\{w_{dji}\}$ of the sentences in a node with theme l as the grouped data. Topic proportions are drawn from $\text{DP}(\alpha_0, \lambda_0)$. Drawing of a topic k for word w_{dji} or w_i depends on the posterior probability

$$p(z_{wi} = k|\mathbf{w}_{dj}, \mathbf{z}_{w(-i)}, c_{dj}, \alpha_0, \lambda_0, \phi_k) \propto p(z_{wi} = k| \mathbf{z}_{w(-i)}, c_{dj}, \alpha_0, \lambda_0)p(w_{dji}|\mathbf{w}_{dj(-i)}, \mathbf{z}_w, \phi_k). \quad (5)$$

Calculating (5) is equivalent to estimating the topic proportion π_{lk}. The first term in RHS of (5) is a distribution over topics derived as an expectation of GEM and is calculated via a product of beta variables using $\text{Beta}\left(\alpha_0\lambda_0, \alpha_0(1-\lambda_0)\right)$. The second term calculates the probability of word w_{dji} given topic model ϕ_k. Given the current status of the sampler, we iteratively sample each variable conditioned on the rest variables. For each document d, the paths c_{dj}, themes l and topics k are sequentially sampled and iteratively employed to update the corresponding posterior probabilities in Gibbs sampling procedure. The true posteriors are approximated by running sufficient iterations of Gibbs sampling. The resulting H2TM is implemented for document summarization.

5. EXPERIMENTS

5.1. Experimental Setup

A series of experiments were conducted to evaluate the proposed H2TM for document summarization. The experiments were performed by using DUC (Document Understanding Conference) 2007 (http://duc.nist.gov/). In DUC 2007, there were 45 super-documents where each document contained 25-50 news articles. The number of total sentences in this dataset was 22961. The vocabulary size was 18696 after removing stop words. This corpus provided the reference summaries, which were manually written for evaluation. The automatic summary for DUC was limited to 250 words at most. The NIST evaluation tool, ROUGE (Recall-Oriented Understudy for Gisting Evaluation), was adopted. ROUGE-1 was used to measure the matched unigrams between reference summary and automatic summary, and ROUGE-L was used to calculate the longest common subsequence between two text datasets. For simplicity, we constrained tree growing to three layers in our experiments. The initial values of three-layer H2TM were specified by $\psi_l = \phi_k = [0.05\, 0.025\, 0.0125]^T$, $\lambda_0 = \lambda_s = 0.35$, $\alpha_s = 100$ and $\gamma_s = 0.5$.

	Recall	Precision	F-measure
H2TM-root	0.4001	0.3771	0.3878
H2TM-leaf	0.4019	0.3930	0.3927
H2TM-MMR	0.4093	0.3861	0.3969
H2TM-path	0.4100	0.3869	0.3976

Table 1. Comparison of recall, precision and F-measure by using H2TM based on four sentence selection methods.

5.2. Evaluation for Summarization

We conduct unsupervised structural learning and provide sentence-level thematic information for document summarization. The thematic sentences are selected from a tree structure where the sentences are allocated in the corresponding tree nodes. The tree model was built by running 40 iterations of Gibbs sampling. The tree path contains sentences for a theme in different layers with varying degree of thematic focus. The thematic sentences are selected by four methods. The first two methods (denoted by H2TM-root and H2TM-leaf) are designed to calculate the Kullback-Leibler (KL) divergence between document model and sentence models using the sentences which are grouped into root node and leaf nodes, respectively. The sentences with small KL divergences are selected. The third method (denoted by H2TM-MMR) is to apply the maximal marginal relevance (MMR) [13] to select sentences from all possible paths. The fourth method (denoted by H2TM-path) chooses the most frequently-visited path of a document among different paths and selects the sentences which are closest to the whole document according to their KL divergence.

Table 1 compares four selection methods to document summarization in terms of recall, precision and F-measure under ROUGE-1. The H2TM-path and H2TM-MMR obtains comparable results. These two methods perform better than H2TM-root and H2TM-leaf. H2TM-path obtains the highest F-measure. The sentences along the most frequently-visited path contain the most representative information for summarization. In the subsequent evaluation, H2TM-path is adopted for comparison with other summarization methods.

Table 2 reports the recall, precision and F-measure of document summarization by using Vector Space Model (VSM), sentence-based LDA [5] and H2TM under ROUGE-1 and ROUGE-L. We also show improvement rates (%) (given in parentheses) of LDA and H2TM over baseline VSM. LDA was implemented for individual sentences by adopting Dirichlet parameter $\alpha = 10$ and fixing the number of topics as 100 and number of themes as 1000. Using LDA, the model size is fixed. This model size is comparable with that of H2TM which is determined autonomously by Bayesian nonparametric learning. The comparison between LDA and H2TM is fair under comparable model complexity. In this evaluation, LDA consistently outperforms baseline VSM in terms of precision, recall and F-measure under different

	ROUGE-1			ROUGE-L		
	Recall	Precision	F-measure	Recall	Precision	F-measure
VSM	0.3262 (-)	0.3373 (-)	0.3310 (-)	0.2971 (-)	0.3070 (-)	0.3013 (-)
LDA	0.3372 (3.4)	0.3844 (14.0)	0.3580 (8.2)	0.2982 (0.4)	0.3395 (10.6)	0.3164 (5.0)
H2TM	0.4100 (25.7)	0.3869 (14.7)	0.3976 (20.1)	0.3695 (24.4)	0.3489 (13.7)	0.3585 (19.0)

Table 2. Comparison of recall, precision and F-measure and their improvement rates (%) over baseline system.

ROUGE measures. Nevertheless, H2TM further improves LDA in presence of different experimental conditions. For the case of ROUGE-1, the improvement rates of F-measure using LDA and H2TM are 8.2% and 20.1%, respectively. The contributions of H2TM come from the flexible model complexity and the structural theme information which are beneficial for document summarization.

6. CONCLUSIONS

This paper addressed a new H2TM for unsupervised learning of latent structure of the grouped data in different levels. A hierarchical theme model was constructed according to a sentence-level nCRP while the topic model was established through a word-level HDP. The snCRP compound HDP was proposed to build H2TM where each theme was characterized by a mixture model of topics. A delicate document representation using the themes in sentence level and the topics in word level was organized. We further presented a TSB process to draw a subtree path for a heterogeneous document and built a hierarchical mixture model of themes according to snCRP. The hierarchical clustering of sentences was realized. The sentences were allocated in tree nodes and the corresponding words in different nodes were drawn by HDP. The proposed H2TM is a general model which can be applied for *unsupervised structural learning* of *different kinds of grouped data*. Experimental results on document summarization showed that H2TM could capture the latent structure from multiple documents and outperform the other methods in terms of recall, precision and F-measure. Further investigations shall be conducted for document classification and information retrieval.

7. REFERENCES

[1] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65, 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 5, pp. 993–1022, 2003.

[3] J.-T. Chien and C.-H. Chueh, "Dirichlet class language models for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 482–495, 2011.

[4] J.-T. Chien and C.-H. Chueh, "Topic-based hierarchical segmentation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 55–66, 2012.

[5] Y.-L. Chang and J.-T. Chien, "Latent Dirichlet learning for document summarization," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1689–1692.

[6] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet process," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[7] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenebaum, "Hierarchical topic models and the nested Chinese restaurant process," in *Advances in Neural Information Processing Systems*, 2004.

[8] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol. 57, no. 2, pp. 7, 2010.

[9] J. Paisley, L. Carin, and D. Blei, "Variational inference for stick-breaking beta process priors," in *Proc. of International Conference on Machine Learning*, 2011.

[10] C. Wang and D. M. Blei, "Variational inference for the nested Chinese restaurant process," in *Advances in Neural Information Processing Systems*, 2009.

[11] J. Pitman, "Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition," *Combinatorics, Probability and Computing*, vol. 11, pp. 501–514, 2002.

[12] R. P. Adams, Z. Ghahramani, and M. I. Jordan, "Tree-structured stick breaking for hierarchical data," in *Advances in Neural Information Processing Systems*, 2010.

[13] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, "Multi-document summarization by sentence extraction," in *Proc. of ANLP/NAACL Workshop on Automatic Summarization*, 2000, pp. 40–48.