Update Summarization Using a Multi-level Hierarchical Dirichlet Process Model

Jiwei Li[1] *Sujian Li*[1] *Xun Wang*[1] *Ye Tian*[1] *Baobao Chang*[1]
(1) Key Laboratory of Computational Linguistics, Peking University, Ministry of Education, CHINA
{bdlijiwei, lisujian, xunwang, ytian, chbb}@pku.edu.cn

ABSTRACT

Update summarization is a new challenge which combines salience ranking with novelty detection. Previous researches usually convert novelty detection to the problem of redundancy removal or salience re-ranking, and seldom explore the birth, splitting, merging and death of aspects for a given topic. In this paper, we borrow the idea of evolutionary clustering and propose a three-level HDP model named h-uHDP, which reveals the diversity and commonality between aspects discovered from two different epochs (i.e. epoch history and epoch update). Specifically, we strengthen modeling the sentence level in the h-uHDP model to adapt to the sentence extraction based framework. Automatic and manual evaluations on TAC data demonstrate the effectiveness of our update summarization algorithm, especially from the novelty criterion.

KEYWORDS : Update summarization, Hierarchical Dirichlet process, Novelty detection.

Proceedings of COLING 2012: Technical Papers, pages 1603–1618,
COLING 2012, Mumbai, December 2012.

1 Introduction

Update summarization aims to generate a short and concise summary for the latest updating topic-related documents (hereafter *update documents* for short), under the assumption that the user has already read the earlier historical documents (*history documents* for short) about the same topic. Recently, there have been many attempts to explore different approaches to generate update summaries. The predominant approaches are mainly built upon the sentence extraction framework.

Update summarization for an evolving topic differs from previous generic summarization for a static topic in that the latter aims to acquire the salient information in one topic, while the former cares for both the salience and the novelty of information. By developing traditional summarization techniques, massive efforts on update summarization have been made to dig out new information (Boudin et al., 2008; Fisher and Boark, 2008; Wan, 2007; Li et al., 2008; Du et al., 2010; Li et al., 2012). The typical examples include the scaled Maximal Marginal Relevance (MMR) algorithm which excludes those sentences similar to the history documents, and some extensions of TextRank such as TimedTextRank (Wan, 2007), PNR^2 (Li et al., 2008), MRSP (Du et al., 2010) which re-rank the salience scores of sentences by employing various kinds of reinforcement between sentences. One problem with these approaches is that they tend to regard update summarization more as a redundancy removal problem than a novelty detection problem. Another problem is that these approaches are mainly based on the computation of lexical similarities between sentences and fail to consider higher level information to avoid semantic redundancy in update summarization.

To solve these two problems, we borrow the techniques of evolutionary clustering which focuses on detecting the dynamics of a given topic. Normally, one topic is described from various specific aspects [1], accompanied with the background information running the whole topic (Chemudugunta et al., 2007; Li et al., 2010). For example, the topic "Quebec independence" may involve the specific aspects including "leader in independence movement", "referendum", "related efforts in independence movement" and so on, while "Quebec" and "independence" are seen as the general background information. The evolving dynamics of a topic is mainly embodied in the birth, splitting, merging and death of the specific aspects (Ren et al., 2008). Then, the commonality and diversity between history documents and update documents can be easily summarized from the aspect level and update summarization is not limited to lexical redundancy removal. Recently, hierarchical Dirichlet process (HDP) (Teh et al., 2006) has been widely used to model the aspects in evolutionary clustering. HDP does not need to predefine the number of clusters, and can be easily and naturally extended to multiple correlated corpora for detecting aspects (Ren et al., 2008; Xu et al., 2008; Zhang et al., 2010; Gao et al., 2011). These distinct advantages make it suitable to update summarization. However, to our best knowledge, no previous work has explored HDP for update summarization.

Aiming at the task of update summarization, in this paper, we develop a novel three-level (i.e. corpus, document set, and document levels) HDP model, called h-uHDP model, which extends the standard HDP to the scenario of two related document sets in different epochs (namely history epoch and update epoch). In h-uHDP, the diversity and connections of aspects between two epochs are naturally modeled: two epochs may share some common aspects; further, some aspects may become outdated while some become popular or some new may appear over time, causing the number of aspects and aspect structures to change at different epochs.

[1] Aspect in this article is usually called *cluster* in evolutionary clustering.

Under the framework of extractive summarization, it is important to acquire the relationship between sentences and aspects for sentence selection. However, in most existing HDP models, the sentence level is disregarded and we cannot directly get the aspect distribution of sentences. Inspired by the progress made in Latent Dirichlet Allocation (LDA) models (Chemudugunta et al., 2007; Li et al., 2010; Delort and Alfonseca, 2012), we newly add the sentence level between the word level and document level in the h-uHDP model. Since neighboring sentences in one document usually talk about one same aspect, we assume that the aspect assignment of each sentence is not conditionally independently. With such assumption, the aspect of each sentence is determined by the aspect distribution of both the document and its neighboring sentences. Our h-uHDP model is capable of mapping multiple levels of information into the latent aspect space.

The rest of this paper is organized as follows. Section 2 discusses the related work on update summarization and evolutionary clustering. Section 3 briefly introduces Dirichlet Process (DP) and Hierarchical Dirichlet Process (HDP). Section 4 presents our proposed aspect model h-uHDP and its inference algorithm. Section 5 addresses the algorithm of update summarization. Section 6 shows the experimental results. Finally, Section 7 concludes the paper.

2 Related work

In this section, we review the related work on update summarization and evolutionary clustering.

2.1 Update summarization

In generic summarization[2], numerous techniques have been developed to measure the salience of sentences and remove the redundancy in summaries, such as the well-known Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998), TextRank (Mihalcea and Tarau, 2004) et al. Some initial work on update summarization inherited the idea of salience ranking in generic summarization and extended the available algorithms to selecting sentences from the newly-coming documents. Boudin et al. (2008) proposed a sentence scoring algorithm derived from MMR and preferred to select those sentences dissimilar to previously read sentences. Fisher and Roark (2008) used a supervised perceptron and simple filtering rules to get the salient sentences for the update documents. Gillick et al. (2008) formulated sentence selection as the problem of integer linear programming and aimed to select a set of sentences that maximize the sum of weights of n-grams covered by the sentences. Adapting the ILP of Gillick et al.(2008), CLASSY by Conroy et al.(2009) seeked to find the sentences that maximize the total approximate oracle scores. Wang and Li (2010) employed an incremental hierarchical clustering algorithm COBWEB to re-organize sentence clusters immediately after new documents/sentences arrive and the most representative sentences for the updated clusters were selected. The graph-based algorithm – TextRank (Mihalcea and Tarau, 2004) has more extensions for update summarization. TimedTextRank by Wan (2007) introduced the time decaying ratio for weighting sentence reinforcement, PNR^2 by Li et al. (2008) added the negative reinforcement between sentences, and MRSP by Du et al. (2010) turned the historical sentences into sink points which are limited their reinforcement with other sentences. Through reinforcement propagation, the salience of sentences in the update documents is influenced by history documents to assure that those sentences with less redundancy with history documents appear in the update summaries. However, they mainly start from the lexical level and cannot explain explicitly what the novel information is.

There are also a few attempts to explore semantic information in update summarization. Steinberger and Jezek (2009) proposed the Iterative Residual Rescaling (IRR) algorithm which

[2] In this paper, generic summarization refers to the non-update summarization.

maps the documents to a set of latent semantic aspects[3]. Then sentences containing novel and significant aspects are then selected for the summary. Inspired by Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Delort and Alfonseca (2012) proposed the DualSum algorithm which designs a nonparametric Bayesian approach to generate four kinds of aspects respectively for background, document, common and novel information. Though these researches have achieved some preliminary findings on exploring semantic information in update summarization, they still cannot present a unified framework to reveal the dynamics of a given topic.

2.2 Evolutional clustering and HDP

Evolutionary clustering is a relatively new research for topic detection, which aims to preserve the smoothness of clustering results over time, while fitting the data of each epoch. The work by Chakrabarti et al. (2006) was probably considered as the first to address the problem of evolutionary clustering. They proposed a general framework of evolutionary clustering and extended two classical clustering algorithms to the evolutionary setting: (1) k-means clustering, and (2) agglomerative hierarchical clustering. Later, Chi et al. (2008) presented two frameworks by incorporating temporal smoothness constraint and applied them on spectral clustering algorithm. While the researches on extending classic clustering algorithms have advanced the literature of evolutionary clustering, they have a very restrictive assumption: the number of clusters over time stays the same. It is clear that this assumption is obviously violated in many real applications.

Recently, HDP has been widely used in evolutionary clustering due to its capability of learning number of clusters automatically and sharing mixture components across different corpora. In HDP, each corpus is modeled by an infinite Dirichlet Process (DP) mixture model, and the infinite set of mixture clusters is shared among all corpora. Sethuraman (1994) gave a stick-breaking constructive definition of DP for arbitrarily measurable base space, which is very useful to model the weight of mixture components in the mixture model. Blackwell and MacQueen (1973) explained DP using the Polya urn scheme, as the predictive distribution of an event is proportional to the frequency of the existing events or to a concentration parameter for an unpresented event. The Polya urn scheme is closely related to the Chinese Restaurant Process (CRP) metaphor, which is applied on HDP demonstrating the 'clustering property' as the 'distribution on partition'. In addition, HDP can also be seen as an LDA-based model, which can automatically and naturally infer the number of clusters from data (Teh et al., 2006). Base on HDP, some algorithms of evolutionary clustering are proposed by incorporating time dependencies, such as DPChain, HDP-EVO, HDP-HMM, dynamic HDP and EvoHDP et al. (Xu et al., 2008; Xu et al., 2008; Ren et al., 2008; Zhang et al., 2010; Gao et al., 2011).

3 DP and HDP

In this section, we briefly introduce Dirichlet Process (DP) and Hierarchical Dirichlet Process (HDP).

A DP can be considered as a distribution of probability measure G. Suppose a finite partition $(T_1,\ldots,T_K)$ in the measure space Θ and a probability distribution G_0 on Θ, we write $G \sim DP(\alpha, G_0)$ if $(G(T_1),\ldots,G(T_K)) \sim Dir(\alpha G_0(T_1),\ldots,\alpha G_0(T_K))$, where α is a positive concentration parameter and G_0 is called a base measure. Sethuraman (1994) showed that a measure G drawn from a DP is discrete by the stick-breaking construction:

[3] aspect is called as topic in the original paper of (Steinberger and Jezek, 2009).

$$\beta_k \sim Beta(1,\alpha),\quad \pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l) = \beta_k(1-\sum_{l=1}^{k-1}\pi_l),\quad \{\phi_k\}_{k=1}^{\infty} \sim G_0,\quad G = \sum_{k=1}^{\infty}\pi_k \delta_{\phi_k} \tag{1}$$

where δ_{ϕ_k} is a probability measure concentrated at ϕ_k. It is important to note that the sequence $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ constructed by Eq. (1) satisfies $\sum_{k=1}^{\infty}\pi_k = 1$ with probability 1. For convenience, we write $\boldsymbol{\pi}$~GEM(α)[4] if $\boldsymbol{\pi}$ is a random probability measure defined by Eq. (1). After observing the draws $\theta_1, \theta_2, ..., \theta_{n-1}$ from *G*, the posterior of *G* still satisfies the *DP* distribution:

$$G \mid \theta_1, \theta_2, ..., \theta_{n-1} \sim DP(\alpha + n - 1, \frac{m_k \delta_{\phi_k} + \alpha G_0}{\alpha + n - 1}) \tag{2}$$

where m_k denotes the number of draws taking the value ϕ_k.

A HDP defines a distribution over a set of DPs. In HDP, a global measure G_0 is distributed as a DP with concentration parameter γ and base measure *H*. Then a set of measures $\{G_j\}_{j=1}^{J}$ is drawn independently from a DP with base measure G_0. Such a process is described as:

$$G_0 \sim DP(\gamma, H), \qquad G_j \mid \alpha_0, G_0 \sim DP(\alpha_0, G_0) \tag{3}$$

For each *j*, let $\{\theta_{ji}\}_{i=1}^{n_j}$ be independent and identically distributed (i.i.d.) random variables drawn from G_j. n_j observations $\{x_{ji}\}_{i=1}^{n_j}$ are drawn from the mixture model:

$$\theta_{ji} \sim G_j, \qquad x_{ji} \sim F(x \mid \theta_{ji}) \tag{4}$$

where $F(x|\theta_{ji})$ denotes the distribution of generating x_{ji}. Equations (3) and (4) complete the definition of a HDP mixture model, whose graphical representation is shown in Figure 1(a).

According to Eq. (1), the stick-breaking construction of HDP can be represented as:

$$\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \sim GEM(\gamma),\quad G_0 = \sum_{k=1}^{\infty}\beta_k \delta_{\phi_k},\quad \boldsymbol{\pi_j} = (\pi_{jk})_{k=1}^{\infty} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}),\quad G_j = \sum_{k=1}^{\infty}\pi_{jk}\delta_{\phi_k} \tag{5}$$

and the corresponding graphical model is shown in Fig. 1(b). We can see that HDP can readily be extended to as many levels as are deemed useful. That is, we can obtain a hierarchy of DPs, where the draw from the DP at a given node serves as a base measure for its children (Teh, 2006).

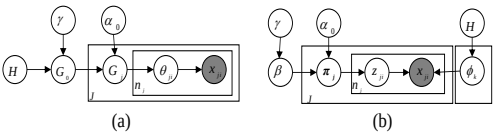

FIGURE 1 – Graphical representation for HDP. (a) original representation. (b) stick-breaking construction

4 h-uHDP model

This section clarifies why and how we propose our improved HDP model (named history-update HDP, *h-uHDP* for short) for the task of update summarization.

In update summarization, a given topic is composed of two document sets (*docset* for short) varying two epochs, namely *history* and *update* epoch. To precisely observe the dynamics of the aspects in one topic, we need to model the aspects over three levels: topic corpus, *docset* at each

[4] GEM stands for Griffiths, Engen, and McCloskey (Teh et al. 2006)

epoch, and document. In such case, we extend the standard HDP to a three level HDP: a set of common aspects on the top level of the hierarchy explicitly address the issue of aspect correspondence between two epochs; the second level is for the aspects at each different epoch, which are considered as the subset of the top level aspects; and the third level is designed for the aspects of each document; the relationship among these three levels of aspects can be obtained through statistical inference. Thus, h-uHDP can naturally model the diversity and connections of aspects between two epochs.

First of all, we introduce some notations in our real data setting of update summarization. We use J^H and J^U to denote the number of documents in the *history* and *update* epochs respectively. For the convenience of description, we use the symbol e in the superscript to denote U or H. Each docset is denoted as $D^e = \{D_j^e\}_{j=1}^{J^e}$ where document D_j^e has N_j^e sentences $\{s_{j,i}^e\}_{i=1}^{N_j^e}$ and the i^{th} sentence in D_j^e has $N_{j,i}^e$ observed word samples $\{x_{j,i,n}^e\}_{n=1}^{N_{j,i}^e}$.

4.1 Model

Our h-uHDP model is an extension of a three-level HDP model which naturally incorporates the levels of corpus, docset and document as shown in Fig. 2. Specifically, we design a two-level HDP respectively for each docset, and these two HDPs share an overall base measure G which is drawn from $DP(\varepsilon,G_0)$ and serves as the overall component bookkeeping for both epochs. We use G^H to denote the global measure for the *history* epoch and call it the *history* global measure. Similarly, G^U is called the *update* global measure. Then, the local measures for each document are denoted as $\{G_j^e\}_{j=1}^{J^e}$, which are drawn from the *history* or *update* global measures. That is, $G_j^e \sim DP(\alpha^e, G^e)$ given $e \in \{H, U\}$.

Then, we introduce the sentence level into the HDP model where each sentence is assigned to one aspect with the consideration of both its neighboring sentences and words contained by this sentence. We use $\theta_{j,i}^e$ to denote the aspect assignment of the i^{th} sentence in D_j^e. There is also a binomial distribution $y \sim binomial(\rho)$, which controls for each sentence how often we encounter a background word, or an aspect word. ρ has a beta prior with parameter β: $\rho \sim beta(\beta)$.

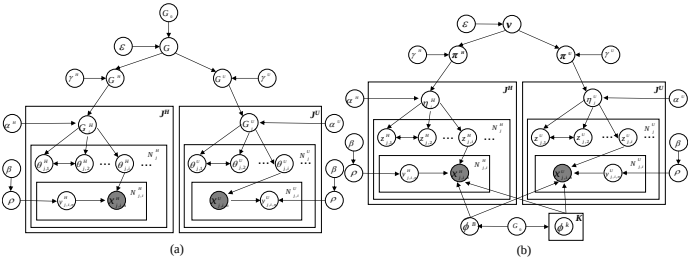

FIGURE 2 –Graphical representation for h-uHDP. (a) original representation. (b) stick-breaking construction

Fig. 2(a) illustrates the graphical representation of h-uHDP model. The generation process of our h-uHDP model is as follows:

1. Draw an overall base measure $G \sim DP(\varepsilon, G_0)$, which denotes the overall aspect distribution for all documents at two epochs.

2. For $e \in \{H, U\}$:

   2.1 Draw the global measure G^e according to the overall measure G. G^e serves as the base measure for each docset. That is, $G^e \sim DP(\gamma^e, G)$.

   2.2 Draw the local measures $\{G_j^e\}_{j=1}^{J^e}$. Each G_j^e for document D_j^e is drawn from the corresponding global measure G^e: $G_j^e \sim DP(\alpha^e, G^e)$.

   2.3 Draw the aspect for sentence $s_{j,i}^e$ according to G_j^e and the aspect assignment of neighboring sentences: $\theta_{j,i}^e \sim g(\theta \mid G_j^e, \boldsymbol{\theta}_{j,-i}^e)$

   2.4 Sample the words $\{x_{j,i,n}^e\}_{n=1}^{N_{j,i}^e}$: $y_{j,i,n}^e \sim binomial(\rho)$, $x_{j,i,n}^e \sim f(x \mid \theta_{j,i}^e, y_{j,i,n}^e)$ where $f(x \mid \theta_{j,i}^e, y_{j,i,n}^e)$ is a distribution parameterized by $\theta_{j,i}^e$ and $y_{j,i,n}^e$.

We can see that the extended three-level HDP model h-uHDP, in fact, considers five levels for a given topic, including word, sentence, document, docset, and corpus. At the same time, aspect assignment dependency between sentences is naturally incorporated in the model.

Next, we will provide the stick-breaking perspective and a Gibbs sampler for model inference.

4.2 The stick-breaking construction

According to the stick-breaking construction of DP, the overall base measure G can be expressed with the following form:

$$G = \sum_{k=1}^{\infty} v_k \delta_{\phi_k}, \quad \boldsymbol{v} \sim GEM(\varepsilon) \tag{6}$$

Then, according to Eq. (5), we can also get the global and local measures with the form as:

$$G^e = \sum_{k=1}^{\infty} \pi_k^e \delta_{\phi_k}, \quad \boldsymbol{\pi}^e \sim DP(\gamma^e, v) \tag{7}$$

and

$$G_j^e = \sum_{k=1}^{\infty} \eta_{jk}^e \delta_{\phi_k}, \ \boldsymbol{\eta}_j^e \sim DP(\alpha^e, \boldsymbol{\pi}^e) \tag{8}$$

As with the standard HDP, we get the stick-breaking construction for h-uHDP, illustrated in Figure 2(b).

Next, we focus on the modeling of the sentence level. $z_{j,i}^e$ is used to indicate the aspect assignment of sentence $s_{j,i}^e$ and the formula of assigning aspect k to $z_{j,i}^e$ is as follows:

$$p(z_{j,i}^e = k \mid \boldsymbol{\phi}, \boldsymbol{z}_{-j,i}^e) \propto f_k(s_{j,i}^e) g(z_{j,i}^e = k \mid \boldsymbol{z}_{-j,i}^e) \tag{9}$$

$$\text{and} \quad g(z_{j,i}^e = k \mid \boldsymbol{z}_{-j,i}^e) = \frac{1}{|N_j^e - 1|} \sum_{d \in [1, N_j^e] - \{i\}} \exp\left\{ \frac{\sigma \cdot \delta(k, z_{j,d}^e)}{|d - i|} \right\} \tag{10}$$

In Formula (9), $f_k(s_{j,i}^e)$ denotes the probability of generating sentence $s_{j,i}^e$ given aspect k and the function $g(\bullet)$ reflects the influence from the neighboring sentences. We use the symbol '-' to denote the exclusion of current sentence or word, and $\boldsymbol{z}_{-j,i}$ means the aspect assignment of all sentences in D_j^e excluding the current sentence. $\delta(k, z_{j,d}^e)$ equals 1 if the aspect assignment of the d^{th} sentence in D_j^e is k, 0 otherwise. The parameter σ (named as *sentence influence factor*) is used to tune the influence from neighboring sentences. Eq. (10) shows that the longer the distance from one sentence to the current sentence, the less the influence that sentence has on the aspect assignment of the current sentence. $y_{j,i,n}^e$ is the indicator variable of word $x_{j,i,n}^e$ and controlled by a binomial distribution with beta prior β. If $y_{j,i,n}^e = 0$, $x_{j,i,n}^e$ is a background word.

If $y^e_{j,i,n} = 1$, $x^e_{j,i,n}$ is an aspect word.

4.3 Inference

For model inference, we use a straightforward Gibbs sampler based on the Chinese Restaurant Franchise (CRF) and the stick-breaking construction. Thus, we begin with an analog of the CRF process for h-uHDP: a document D^e_j corresponds to a *restaurant*, and a sentence $s^e_{j,i}$ corresponds to a customer. Different from the standard HDP, one customer in our model is seen as a family which includes a few persons. Here, we assume that the persons in one family usually have the same preference for one dish at one table except some persons shown no preference for any food. The general *background dish* is assigned to the persons having no preference and a specific dish k is assigned to those persons having preference. The global menu of dishes is denoted by K+1 i.i.d. random variables $\phi_B, \phi_1, \ldots, \phi_K$ distributed according to G_0. We also introduce variables, ψ^e_{jt}, to represent the dish served at table t in restaurant j. To denote the associations among $\theta^e_{j,i}$, ψ^e_{jt} and ϕ_k, we let t^e_{ji} be the index of ψ^e_{jt} associated with $\theta^e_{j,i}$, and let k^e_{jt} be the index of ϕ_k associated with ψ^e_{jt}. In the CRF metaphor, customer $s^e_{j,i}$ sits at table t^e_{ji} while table t in restaurant D^e_j serves dish k^e_{jt}.

We also need a notation for counts. Specifically, we need to record the counts of families, persons and tables. Marginal counts are represented with dots in the subscript. $n^e_{jt\bullet}$ represents the number of families in restaurant j at table t in the corresponding epoch, and $n^e_{j\bullet k}$ represents the number of families in restaurant j eating dish k in the corresponding epoch. The notation m^e_{jk} denotes the number of tables in restaurant j serving dish k in one epoch, $m^e_{j\bullet}$ denotes the number of tables in restaurant j in one epoch, $m^e_{\bullet k}$ denotes the number of tables serving dish k in one epoch, and $m^e_{\bullet\bullet}$ denotes the total number of tables in each epoch. The notation above with removing the superscript represents the corresponding counts in both epochs. For example, $n_{\bullet\bullet k}$ is the total number of customers assigned to aspect k in both epochs and $m_{\bullet k}$ denotes the total number of tables serving dish k in both epochs.

In our implementation, we first sample the index variables t^e_{ji} and k^e_{jt}. Then the $\theta^e_{j,i}$ and ψ^e_{jt} can be reconstructed from their index variables and ϕ_k, which makes the MCMC sampling more efficient (Blei et al., 2006).

Sampling ***t****.* Due to the space limit, we would just show the sampling formula without derivation. The likelihood due to $s^e_{j,i}$ given $t^e_{ji} = t$ for some previously t is $f^{-s^e_{j,i}}_{k^e_{jt}}(s^e_{j,i})$. The likelihood of generating $s^e_{j,i}$ given $t^e_{ji} = t^{new}$ can be calculated by integrating out the possible values of $k_{jt^{new}}$. K^e denotes the set of aspects assigned in current epoch. The prior probability that t^e_{ji} takes on a previously used t is calculated according to $g(z^e_{j,i} \mid \mathbf{z}^e_{\mathbf{-j,i}})$, $f^{-s_{ji}}_{k_{jt}}(s^e_{j,i})$ and $n^e_{jt\bullet}$, whereas the probability that it takes on a new value (i.e. $t^{new} = m^e_{j\bullet} + 1$) is proportional to α^e. Then, the conditional distribution of t^e_{ji} given the rest of the variables is:

$$p(t^e_{ji} = t \mid \mathbf{t}_{-ji}, \mathbf{k}) \propto \begin{cases} g(z^e_{j,i} = k \mid \mathbf{z}^e_{\mathbf{-j,i}}) n^e_{jt\bullet} f^{-s_{ji}}_{k_{jt}}(s^e_{j,i}) & \textbf{\textit{if } } t \textbf{\textit{ previously used}} \\ \alpha^e P(s^e_{j,i} \mid \mathbf{t}_{-ji}, t^e_{ji} = t^{new}, \mathbf{k}) & \textbf{\textit{if }} t = t^{new} \end{cases} \tag{11}$$

and
$$P(s^e_{j,i} \mid \boldsymbol{t}_{-ji}, t_{ji} = t^{new}, \boldsymbol{k}) = \sum_{k \in K^e} g(z^e_{j,i} = k \mid \boldsymbol{z}^e_{-j,i})\left(\frac{m^e_{\bullet k}}{m^e_{\bullet\bullet} + \gamma^e} + \frac{\gamma^e}{m^e_{\bullet\bullet} + \gamma^e} \cdot \frac{m_{\bullet k}}{m_{\bullet\bullet} + \varepsilon}\right) f_k^{-s^e_{j,i}}(s^e_{j,i}) + \sum_{k \in K^{\{H,U\}} - K^e} \frac{\gamma^e}{m^e_{\bullet\bullet} + \gamma^e} \cdot \frac{m_{\bullet k}}{m_{\bullet\bullet} + \varepsilon} f_k^{-s^e_{j,i}}(s^e_{j,i}) + \frac{\gamma^e}{m^e_{\bullet\bullet} + \gamma^e} \cdot \frac{\varepsilon}{m_{\bullet\bullet} + \varepsilon} f_{k^{new}}^{-s_{j,i}}(s^e_{j,i}) \tag{12}$$

where $m^e_{\bullet\bullet}$ denotes the total number of tables in epoch *e* and $K^{\{H,U\}}$ means the set of the aspects available in the two epochs.

According to the distribution $y^e_{j,i,n} \sim binomial(\rho)$ and $\rho \sim beta(\beta)$, we can get the conditional probability of generating sentence $s^e_{j,i}$ given a specified aspect *k*:

$$f_k^{-s^e_{j,i}}(s^e_{j,i}) = \begin{cases} \dfrac{\Gamma(n_{\bullet\bullet k} + V\beta)}{\Gamma(n_{\bullet\bullet k} + \sum_k A^{(k)}_{(s^e_{j,i})} + V\beta)} \prod_{\substack{x^e_{j,i,n} = x^e_{j,i} \\ y^e_{j,i,n} = 1}} \dfrac{\Gamma(E^{(k)}_{(x^e_{j,i,n})} + E^{(s^e_{j,i})}_{(x^e_{j,i,n})} + \beta)}{\Gamma(E^{(k)}_{(x^e_{j,i,n})} + \beta)} & \text{if } k \text{ previously used} \\ \dfrac{\Gamma(V\beta)}{\Gamma(\sum_k A^{(k)}_{(s^e_{j,i})} + V\beta)} \prod_{\substack{x^e_{j,i,n} = x^e_{j,i} \\ y^e_{j,i,n} = 1}} \dfrac{\Gamma(E^{(s^e_{j,i})}_{(x^e_{j,i,n})} + \beta)}{\Gamma(\beta)} & k = k^{new} \end{cases} \tag{13}$$

$A^{(k)}_{(s^e_{j,i})}$ denotes the number of words that belong to aspect *k* in sentence $s^e_{j,i}$. It is obvious that background words do not influence the aspect assignment of sentences. $n_{\bullet\bullet k}$ is the total number of words assigned to aspect *k* in both epochs. *V* represents the size of vocabulary. $E^{(k)}_{(x^e_{j,i,n})}$ denotes the total number of times that word $x^e_{j,i,n}$ belongs to topic *k* in both docsets, and $E^{(s^e_{j,i})}_{(x^e_{j,i,n})}$ denotes the number of times that word $x^e_{j,i,n}$ exists in $s^e_{j,i}$.

If the sampled value of t^e_{ji} is t^{new}, then we can sample $k^e_{jt^{new}}$ according to (12):

$$p(k^e_{jt^{new}} = k \mid \boldsymbol{t}, \boldsymbol{k}^e_{-jt^{new}}) \propto \begin{cases} g(z^e_{j,i} = k \mid \boldsymbol{z}^e_{-j,i})\left(\dfrac{m^e_{\bullet k}}{m^e_{\bullet\bullet} + \gamma^e} + \dfrac{\gamma^e}{m^e_{\bullet\bullet} + \gamma^e} \cdot \dfrac{m_{\bullet k}}{m_{\bullet\bullet} + \varepsilon}\right) f_k^{-s^e_{j,i}}(s^e_{j,i}) & \textbf{if } k \in K^e \\ \dfrac{\gamma^e}{m^e_{\bullet\bullet} + \gamma^e} \cdot \dfrac{m_{\bullet k}}{m_{\bullet\bullet} + \varepsilon} f_k^{-s^e_{j,i}}(s^e_{j,i}) & \textbf{if } k \in K - K^e \\ \dfrac{\gamma^e}{m^e_{\bullet\bullet} + \gamma^e} \dfrac{\varepsilon}{m_{\bullet\bullet} + \varepsilon} f_{k^{new}}^{-s^e_{j,i}}(s^e_{j,i}) & \textbf{if } k = k^{new} \end{cases} \tag{14}$$

All the counts above except $A^{(k)}_{(s^e_{j,i})}$ and $E^{(s^e_{j,i})}_{(x^e_{j,i,n})}$ exclude the current sentence.

Sampling $\boldsymbol{k}$. Because the process of sampling $\boldsymbol{t}$ actually changes the component member of tables, we continue to sample k^e_{jt} for each table. The conditional probability $p(k^e_{jt} = k \mid \boldsymbol{t}, \boldsymbol{k}_{-jt})$ can be calculated similar to Eq. (14) and it is noted that a set of customers (not one custom) at table *t* should be considered.

Sampling $\boldsymbol{y}$. $y^e_{j,i,n}$ determines whether $x^e_{j,i,n}$ is a background word or an aspect word. If $y^e_{j,i,n} = 0$, $x^e_{j,i,n}$ is a background word, otherwise assigned to aspect *k*. we sample $y^e_{j,i,n}$ as:

$$p(y^e_{j,i,n} \mid z^e_{j,i} = k, \boldsymbol{y}_{-i,j,n}) = \begin{cases} \dfrac{C_{(0)} + \beta}{C_{(\bullet)} + 2\beta} \cdot \dfrac{C^{(x^e_{j,i,n})}_{(0)} + \lambda}{C_{(0)} + V\lambda} & \textbf{if } y^e_{j,i,n} = 0 \\ \dfrac{C_{(1)} + \beta}{C_{(\bullet)} + 2\beta} \cdot \dfrac{C^{(x^e_{j,i,n})}_{(k)} + \lambda}{C^{(k)}_{(1)} + V\lambda} & \textbf{if } y^e_{j,i,n} = 1 \end{cases} \tag{15}$$

where $C_{(\bullet)}$ denotes the total number of words in both docsets, $C_{(0)}$ denotes the total number of background words, $C_{(1)}$ denotes the total number of aspect words, and $C^{(k)}_{(1)}$ denotes the total

number of words that are assigned to aspect k. $C_{(0)}^{(x_{j,i,n}^e)}$ represents the number of times that word $x_{j,i,n}^e$ is assigned to background word and $C_{(k)}^{(x_{j,i,n}^e)}$ represents the total number of times that word $x_{j,i,n}^e$ is assigned to aspect k. The base measure G_0 was set a symmetric Dirichlet distribution with parameters λ (e.g. 0.5).

Based on Equations (11), (12) and (14), the aspect assignment probability of each sentence can be calculated as:

$$p(s_{j,i}^e = k \mid \boldsymbol{k}_{-ji}^e) \propto \begin{cases} g(z_{j,i}^e = k \mid \boldsymbol{z}_{-j,i}^e) n_{j\bullet k}^e f_k^{-s_{j,i}^e}(s_{j,i}^e) + \alpha^e g(z_{j,i}^e = k \mid \boldsymbol{z}_{j,i}^e) & \\ \times \left(\frac{m_{\bullet k}^e}{m_{\bullet\bullet}^e + \gamma^e} + \frac{\gamma^e}{m_{\bullet\bullet}^e + \gamma^e} \frac{m_{\bullet k}}{m_{\bullet\bullet} + \varepsilon}\right) f_k^{-s_{j,i}^e}(s_{j,i}^e) & \text{if } k \in K^e \\ \alpha^e \frac{\gamma^e}{m_{\bullet\bullet}^e + \gamma^e} \frac{m_{\bullet k}}{m_{\bullet\bullet} + \varepsilon} f_k^{-s_{j,i}^e}(s_{j,i}^e) & \text{if } k \in K^{(H,U)} - K^e \\ \alpha^e \frac{\gamma^e}{m_{\bullet\bullet}^e + \gamma^e} \frac{\varepsilon}{m_{\bullet\bullet} + \varepsilon} f_{k^{new}}^{-s_{j,i}^e}(s_{j,i}^e) & \text{if } k = k^{new} \end{cases} \tag{16}$$

As for the concentration parameters of h-uHDP, i.e., ε , α^e and γ^e, we sample them from a vague gamma prior which is set to be $Ga(10.0, 1.0)$. The sampling method is the same as that in (Teh et al., 2006).

5 Update Summarization with h-uHDP model

The task of update summarization aims to produce an update summary for the documents in the *update* epoch, assuming that users already read earlier documents in the *history* epoch. That is, we need to boost sentences in *update* epoch that can bring out important and novel information. On one hand, the generated summary should extract the main content in D^U, and on the other hand, the summary should avoid mentioning too much old information in D^H. To care for these two points, we propose a sentence selection strategy based on Kullback-Leibler (KL) divergence, which has been widely used in extractive summarization (Haghighi and Vanderwende, 2009; Mason and Charniak, 2011; Delort and Alfonseca, 2012).

Given the *history* sentence set S^H and the *update* sentence set S^U, we propose a function to score a set of sentences *Sum* which is a subset S^U.

$$Score(Sum) = KL(p_{S^H} \mid\mid p_{Sum}) - \kappa KL(p_{S^U} \mid\mid p_{Sum}) \tag{17}$$

In the equation, the first term means the prize on the divergence from epoch *history* and the second term represents the penalty on the divergence from epoch *update*. The parameter κ (called as *epoch balance factor*) is used to tune the weights of two KL distances. p_{Sum} is the empirical aspect distribution of the candidate summary *Sum*. p_{S^H} and p_{S^U} respectively denote the aspect distribution of S^H and S^U. $KL(p_{S^e} \mid\mid p_{Sum})$ $(e \in \{H, U\})$ represents the KL divergence given by $\sum_{k=1}^{K} p(S^e \mid k) \log \frac{p(S^e \mid k)}{p(Sum \mid k)}$. $p(\cdot \mid k)$ represents the probability distribution of a set of sentences on a specific aspect k, and is calculated based on the aspect assignment probability of each sentence which can be obtained according to Eq. (16).

$$p(S^e \mid k) = \frac{1}{\sum_j N_j^e} \sum_{s \in S^e} p(s = k), \quad p(Sum \mid k) = \frac{1}{\mid Sum \mid} \sum_{s \in Sum} p(s = k) \tag{18}$$

Generally, an optimum update summary should have the aspect distribution which approximates to p_{S^U} as possible and keep far away from the distribution p_{S^H}. Let Sum^* denote the optimum update summary. We can get Sum^* that maximizes the scoring function.

$$Sum^* = \underset{Sum \subseteq S^U \,\&\&\, words(Sum) \le L}{\arg\max} Score(Sum) \tag{19}$$

Since the problem of finding the subset of sentences from a collection that maximize the scoring function is NP-complete, a greedy algorithm is applied by adding sentences one by one. We use *Y* to denote the sentence set which contains the selected summary sentences. The algorithm first initializes *Y* to ϕ and *X* to S^U. During each iteration, we select from *X* one sentence (i.e. s_m) which makes $Score(s_m \cup Y)$ have the highest score. To avoid aspect redundancy in the summary, we also adopt the MMR strategy in the process of sentence selection. That is, for each s_m, we compute the semantic similarity between s_m and each sentence s_t in set *Y* as follows:

$$\cos\_sem(s_m, s_t) = \frac{\sum_k p(s_m \mid k) \cdot p(s_t \mid k)}{\sqrt{\sum_{k=1}^{K} p^2(s_m \mid k)} \cdot \sqrt{\sum_{k=1}^{K} p^2(s_t \mid k)}} \tag{20}$$

6 Experiments

In our experiments, we use four years of TAC (2008-2011) data, which contain 44-48 topics per year. For each topic, two docsets (named docset *H* and *U*) are given to respectively describe the *history* epoch and the *Update* epoch. Table 1 illustrates the number of topics, averaged number of documents per docset, and averaged number of sentences per docset for each year's data.

TAC	**2008**		**2009**		**2010**		**2011**	
Topics #	48		44		46		44	
docset	***H***	***U***	***H***	***U***	***H***	***U***	***H***	***U***
Avg Doc # per docset	10	10	10	10	10	10	10	10
Avg Sen # per docset	236.5	222.4	253.5	228.3	238.6	230.2	208.9	210.5

TABLE 1 – Experiment data (TAC 2008 - 2011).

As for the automatic evaluation of summarization, we still use the widely used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2003) measures, including ROUGE-1, ROUGE-2, and ROUGE-SU4[5] and their corresponding 95% confidential intervals. In order to obtain a more comprehensive measure of summary quality, we also conduct manual evaluation on TAC 2011 dataset with the reference to (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2011; Delort and Alfonseca, 2011).

6.1 Parameter tuning

To get the final update summary using the h-uHDPSum algorithm, we still need to determine two parameters: sentence influence factor σ in Eq. (10) and epoch balance factor κ in Eq. (17). The combination of the two factors makes it hard to find a global optimized solution. So we apply a

[5]Jackknife scoring for ROUGE is used in order to compare with the human summaries.

gradient search strategy. At first, the epoch balance factor κ is fixed to a given value. Then the performance using different values of σ is evaluated. After that, we fix σ with the value which has achieved the best performance, and conduct experiments to find an appropriate value for κ. TAC 2008 and 2009 datasets are used as training data to tune these two parameters.

Firstly, κ is set to the value of 1, i.e. the prize on the divergence from epoch *history* is as important as the penalty on the divergence from epoch *update*. Reviewing Eq. (10), we can see that, the aspect assignment of one sentence is mainly determined by its neighboring sentences when σ is set a large value, whereas the influence from other sentences is not considered at all when σ is set 0. In the first place, we experiment the h-uHDPsum algorithm by setting σ in the range from 0 to 10 with interval of 1. The ROUGE scores drop sharply when σ is set a value larger than 2.0. Next, σ is set in the range from 0.0 to 2.0 with interval of 1.0. Fig. 4 presents the ROUGE-2 and ROUGE-SU4 evaluation results of h-uHDPSum, with regard to different values of σ. We find that the ROUGE scores reach their peak at around 1.0 and drop afterwords. The experimental results conform to our expectation and verify that the h-uHDP model is reasonable by considering the influence among sentences.

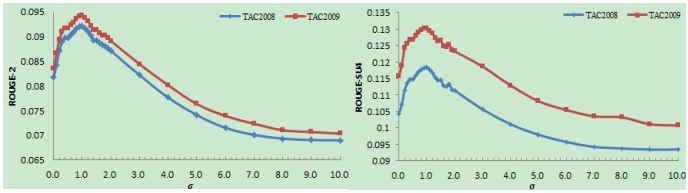

FIGURE 4 – Tuning parameter σ when κ is set to 1.

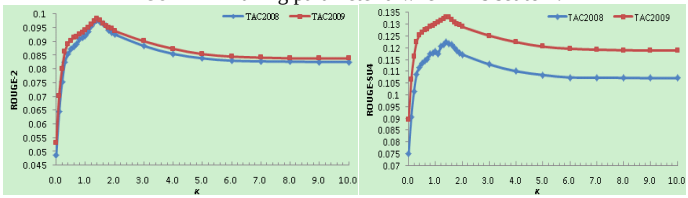

FIGURE 5 – Tuning parameter κ when σ is set to 1.

Next, we fix the sentence influence factor σ at 1.0 and tune the parameter κ. From Eq. (17), we can see that κ is used to balance the prize for the divergence from *history* epoch and penalty on the divergence from the update epoch. When κ is set as the value of 0, the scoring of sentences is only determined by docset *H*. That is, a sentence is likely to be selected into summary, only when it has a large divergence of aspect distribution from docset *H*. When κ is set a larger value, penalty on the divergence from docset *U* is more considered. Similar to the process of tuning σ, the performance using different values of κ ranging from 0 to 10 with interval of 1 is evaluated. We find that the peak performance of κ should be located in the range of [0.0, 2.0]. Thus, we conduct experiments to find an appropriate value for κ in the range from 0.0 to 2.0 with interval of 0.1. Fig. 5 shows the performance of h-uHDPSum with respect to κ. Performance gets better as κ increases from 0 to 1.4, and then declines gently until κ arrives at 3.0. Afterwards, the curve becomes smooth and means that the summarization algorithm is mainly up to docset *U* to decide which sentences to select. Parameters σ and κ are respectively set as 1.0 and 1.4 in the h-uHDPSum algorithm.

6.2 Comparison with other approaches

In this subsection, we compare our h-uHDPSum algorithm with several baseline methods on TAC 2010 and TAC 2011 datasets. One kind of baseline methods consists of the top three performing systems (denoted as SysRank 1st, 2nd and 3rd) on update summarization tasks according to the ROUGE-2 metric on TAC2010 and TAC2011. From Table 3, we can see that our approach obviously outperformed the top three participating systems both on TAC2010 and TAC2011, with respect to the ROUGE-2 and ROUGE-SU4 scores along with the corresponding 95% confidence intervals.

	TAC2010		TAC2011	
	ROUGE-2	**ROUGE-SU4**	**ROUGE-2**	**ROUGE-SU4**
h-uHDPSum	**0.0857(0.0784-0.0930)**	**0.1255(0.1182-0.1328)**	**0.1017(0.0910-0.1034)**	**0.1364(0.1265-0.1473)**
SysRank 1st	0.0799(0.0747-0.0851)	0.1198(0.1154-0.1244)	0.0959(0.0894-0.1029)	0.1309(0.1251-0.1366)
SysRank 2nd	0.0790(0.0740-0.0842)	0.1187(0.1142-0.1234)	0.0924(0.0857-0.0993)	0.1274(0.1217-0.1334)
SysRank 3rd	0.0729(0.0682-0.0779)	0.1080(0.1041-0.1122)	0.0863(0.0808-0.0920)	0.1280(0.1229-0.1330)
h-uHDPSum-noBG	0.0812(0.0767-0.0852)	0.1199(0.1120-0.1278)	0.0931(0.0874-0.0988)	0.1310(0.1258-0.1362)
2LevHDPSum	0.0780(0.0709-0.0851)	0.1171(0.1110-0.1232)	0.0917(0.0863-0.0971)	0.1315(0.1227-0.1401)
HDPSum	0.0708(0.0631-0.0785)	0.1091(0.1014-0.1176)	0.0842(0.0782-0.0902)	0.1218(0.1163-0.1273)
2LevLDASum	0.0720(0.0687-0.0793)	0.1152(0.1067-0.1237)	0.0879(0.0831-0.0925)	0.1310(0.1255-0.1366)
LDASum	0.0649(0.0594-0.0704)	0.1027(0.0936-0.1118)	0.0767(0.0704-0.0830)	0.1074(0.1015-0.1134)

TABLE 2 – Performance Comparison on TAC2010 and TAC2011.

To illustrate the effectiveness of our aspect modeling technique, we provide five other baseline systems which adopt different aspect modeling techniques. The systems *h-uHDPSum-noBG* and *2LevHDPSum* can be seen the simplified versions of *h-uHDPSum*. *h-uHDPSum-noBG* is the same as *h-uHDPSum* except that the general background information is not considered, whereas *2LevHDPSum* is a two-level (i.e. document level and sentence level) HDP model where the docset level is removed. At the same time, we implement one standard HDP model for comparison. As shown in Table 2, *h-uHDPSum* is better than both *h-uHDPSum-noBG*, *2LevHDPSum* and *HDPSum*, which verifies that the identification of background words or the introduction of the docset level can promote the performance of update summarization. Even without consideration of the background information, we can see the *h-uHDPSum-noBG* approach can be comparable to the best participating system of TAC evaluations. In addition, to compare with another popular modeling technique - Latent Dirichlet Allocation (LDA), we design a two-level LDA-based system *2LevLDASum* and a standard LDA-based system *LDASum*. Both *2LevLDASum* and *2LevHDPSum* are similar as possible beyond the distinction that *2LevLDASum* assume a fixed finite number of aspects[6] while *2LevHDPSum* does not. We can see that *2LevHDPSum* is better than *2LevLDASum* and *HDPSum* better than *LDASum* in performance. This can be easily explained, novel aspects can be automatically detected and the aspect number is determined naturally in HDP-based models. In contrast, how to determine the aspect number in the LDA-based models is still an open problem. This is also the reason why we select HDP as the foundation of our aspect modeling technique.

6.3 Manual evaluation

In order to obtain a more accurate measure of summary quality, manual evaluation is required. In this section, we compare our *h-uHDPSum* approach with *2LevLDASum* and the best participating

[6] In our experiments, the aspect number is set as 10, 20, 30 and 40 respectively and we select the best performed result with the aspect number as 20.

system (*Peer 43*). Similar to the manual evaluation in TAC, human assessors assign a score to each summary with respect to each of the following four criteria: 1) Overall responsiveness (overall performance in terms of content and fluency), 2) Focus (containing less irrelevant details), 3) Novelty (containing novel information beyond docset *H*), 4) Non-redundancy (repeating less the same information). The score is an integer between 1 (very poor) and 5 (very good). We randomly select 28 topics from TAC 2011 data and assign each topic to three different assessors[7]. In Table 3, the left four columns report the average scores of each criterion for the three systems. The experimental results indicate that *h-uHDPSum* is significantly better than both *Peer 43* and *2LevLDASum* (based on paired t-test with p-value < 0.01).

Simultaneously, a fairly standard approach for manual evaluation is conducted through pairwise comparison (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2011). According to the rating scores, each pair of summaries is judged which one is better under each criterion. If two summaries have the same score, they are judged a tie (of the equal quality). We record the times of 'winning' (having a higher score) and tie for each system. In Table 3, the right six columns show the evaluation results in frequencies respectively for *h-uHDPSum* vs. *Peer 43*, and *h-uHDPSum* vs. *2LevLDASum*. The experimental results also indicate that *h-uHDPSum* is significantly better than both *Peer 43* and *2LevLDASum*. We also observe that the winning times of *h-uHDPSum* under the novelty criterion is much more than those under the other criteria. This indicates that our approach can exhibit a clear advantage of promoting the novelty performance in update summaries.

				h-uHDPSum vs. *Peer 43*			*h-uHDPSum* vs. *2LevLDASum*		
	h-uHDPSum	*Peer 43*	*2LevLDASum*	*h-uHDP.*	Tie	*Peer 43*	*h-uHDP.*	Tie	*2LevLDA.*
Overall	**3. 94**	3. 65	3. 56	**31**	39	14	**50**	19	15
Focus	**4. 06**	3. 75	3. 65	**36**	26	22	**33**	48	3
Novelty	**4. 17**	3. 69	3. 67	**52**	13	19	**62**	6	16
Non-redund.	**4. 18**	3. 91	3. 98	**41**	22	21	**36**	44	4

TABLE 3 – Results of manual evaluation on TAC2011.

Conclusion

In this paper, we propose a novel approach based on a three-level HDP model h-uHDP for update summarization. The h-uHDP model can detect the birth, splitting, merging and death of specific aspects and the general background information for a given topic. Under the sentence extraction based framework of summarization, we especially strengthen modeling the sentence level in h-uHDP, where the aspect assignment of each sentence is influenced by its neighboring sentences. Based on h-uHDP, we propose a sentence selection strategy adopting KL divergence, which cares for both salience and novelty of sentences. Automatic and manual evaluations on TAC data illustrate that our approach obviously outperforms the state-of-the-art approaches.

Acknowledgments

The research work described in this paper has been partially supported by NSFC grants (No.90920011 and No.61273278), NSSFC grant (No: 10CYY023), National Key Technology R&D Program (No: 2011BAH10B04-03), and National High Technology R&D Program (No. 2012AA011101). We also thank the three anonymous reviewers for their helpful comments. Corresponding authors: Sujian Li (lisujian@pku.edu.cn) and Baobao Chang (chbb@pku.edu.cn).

[7] Each topic includes three update summaries generated by the three systems. A total of 1008 (28topics * 3systems*3persons*4criteria) scores need to be ranked. Six assessors participate the scoring.

References

Blackwell, D. and MacQueen J. B. (1973). Ferguson distributions via Polya urn schemes. The Annals of statistics, 1: 353-355.

Blei, D., Andrew, Y. N. and Jordan, M. L. (2003). Latent Dirichlet Allocation, in *Journal of Machine Learning Research*, vol 3: 993-1022.

Boudin, F., El-Beze, M. and Torres-Moreno, J. (2008). A scalable MMR approach to sentence scoring for multi-document update summarization. In *COLING 2008*, volume: Posters, pages 23–26.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st SIGIR*, pages 335-336.

Celikyilmaz, A. and Hakkani-Tur, D. (2011). Discovery of topically coherent sentences for extractive summarization. In *Proc. of the 49th ACL*, pages 491-499.

Chakrabarti, D., Kumar, R. and Tomkins, A. (2006). Evolutionary clustering, in *Proc. of KDD2006*, pages 554‑560.

Chemudugunta, C., Smyth, P. and Steyvers, M. (2007). Modeling general and specific aspects of documents with a probabilistic topic model. In *Neural Information Processing Systems*. pages 241-248.

Chi, Y., Song, X., Zhou, D., Hino, K. and Tseng, B. L. (2007). Evolutionary spectral clustering by incorporating temporal smoothness, In *Proc. of KDD 2007*, pages 153‑162.

Delort, J. and Alfonseca, E. (2012). DualSum: a topic-Model based approach for update summarization, In *Proc. of the 13th EACL*, pages 214-223.

Du, P., Guo, J., Zhang, J. and Cheng, X. (2010). Manifold ranking with sink points for update summarization. In *Proc. of CIKM'2010*, pages 1757-1760.

John, C. M., Judith, S. D. and Dianne, O. P. (2009). Summarization and metrics. In *Proc. of TAC'09.*

Fisher, S. and Roark, B. (2008). Query-focused supervised sentence ranking for update summaries. In *Proc. of 1st Text Analysis Conference, TAC-2008*.

Gao, Z. J., Song, Y. and Liu, S. (2011). Tracking and connecting topics via incremental hierarchical Dirichlet Processes, In *ICDE 2011*. pages 1056-1061.

Gillick, D., Favre, B., and Hakkani-Tur, D. (2008). The ICSI summarization system at TAC 2008. In *Proc. Of TAC'08.*

Griffiths, T. L., Steyvers, M., Blei, D. M. and Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Neural Information Processing Systems*, vol 17, pages 537-544.

Gruber, A., Weiss, Y. and Rosen-Zvi, M. (2007). Hidden topic Markov models. In *Proc. of the conference on Artificial Intelligence and Statistics.*

Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages. 362-370.

Li, P., Jiang, J and Wang, Y. (2010). Generating templates of entity summaries with an entity-aspect model and pattern mining, in *ACL 2010*, pages. 640-649.

Li, W., Wei, F., Lu, Q. and He, Y. (2008). *PNR*^{2}*:* ranking sentences with positive and negative reinforcement for query-oriented update summarization. *In Proc. of COLING '08*. Vol 1, pages 489–496.

Li, X., Du, L., and Shen, Y. (2012). Update summarization via graph-based sentence ranking, *IEEE Transactions on Knowledge and Data Engineering*.

Lin, C. Y. and Hovy, E. H. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. *In Proc. of HLT-NAACL 200*3, pages 71-78.

Mason, R. and Charniak, E. (2011). Extractive multi-document summaries should explicitly not contain document-specific content. In *Proc. of WASDGML'11*, pages 49-54.

Mihalcea, R. and Tarau, P. (2004). TextRank: bringing order into texts. In *Proc. of EMNLP '04.*

Ren, L., Dunson, D. B. and Carin, L. (2008). The dynamic hierarchical Dirichlet process. In *ICML' 08*, pages: 824- 831.

Sethuraman J. (1994). A constructive definition of Dirichlet priors. In *Statistica Sinaca*, Vol 2 pages: 639-650.

Shen, C. and Li, T. (2010). Multi-document summarization via the minimum dominating set. *In COLING '10*, pages 984–992.

Steinberger, J. and Jezek, K. (2009). Update summarization based on novel topic distribution. *In Proceedings of DocEng'09*. pages 205-213.

Teh, Y., Jordan, M., Beal, M., and D. Blei. (2006). Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, vol. 101, pages 1566‑1581.

Wan, X. (2007). TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization. *In Proc. of SIGIR'07*, pages 867-868.

Wang, D. and Li, T. (2010). Document update summarization using incremental hierarchical clustering. In *CIKM'10*, pages 279 ‑ 288.

Xu, T., Zhang, Z. M., Yu, P. S. and Long. B. (2008). Dirichlet process based evolutionary clustering. In *ICDM'08*. pages 648-657.

Xu, T., Zhang, Z. M., Yu, P. S., and Long. B. (2008). Evolutionary clustering by hierarchical Dirichlet process with hidden Markov state. In *ICDM'08*. pages 658-667.

Zhang, J., Song, Y., Zhang, C. and Liu, S. (2010). Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora, in *KDD* 2010. pages: 1079‑1088.