Evolutionary Hierarchical Dirichlet Process for Timeline Summarization

Jiwei Li
School of Computer Science
Cornell University
Ithaca, NY, 14853
jl3226@cornell.edu

Sujian Li
Laboratory of Computational Linguistics
Peking University
Beijing, P.R.China, 150001
lisujian@pku.edu.cn

Abstract
Timeline summarization aims at generating concise summaries and giving readers a faster and better access to understand the evolution of news. It is a new challenge which combines salience ranking problem with novelty detection. Previous researches in this field seldom explore the evolutionary pattern of topics such as birth, splitting, merging, developing and death. In this paper, we develop a novel model called Evolutionary Hierarchical Dirichlet Process (EHDP) to capture the topic evolution pattern in timeline summarization. In EHDP, time varying information is formulated as a series of HDPs by considering time-dependent information. Experiments on 6 different datasets which contain 3156 documents demonstrates the good performance of our system with regard to ROUGE scores.

1 Introduction
Faced with thousands of news articles, people usually try to ask the general aspects such as the beginning, the evolutionary pattern and the end. General search engines simply return the top ranking articles according to query relevance and fail to trace how a specific event goes. Timeline summarization, which aims at generating a series of concise summaries for news collection published at different epochs can give readers a faster and better access to understand the evolution of news.

The key of timeline summarization is how to select sentences which can tell readers the evolutionary pattern of topics in the event. It is very common that the themes of a corpus evolve over time, and topics of adjacent epochs usually exhibit strong correlations. Thus, it is important to model topics across different documents and over different time periods to detect how the events evolve.

The task of timeline summarization is firstly proposed by Allan et al. (2001) by extracting clusters of noun phases and name entities. Chieu et al. (2004) built a similar system in unit of sentences with interest and burstiness. However, these methods seldom explored the evolutionary characteristics of news. Recently, Yan et al. (2011) extended the graph based sentence ranking algorithm used in traditional multi-document summarization (MDS) to timeline generation by projecting sentences from different time into one plane. They further explored the timeline task from the optimization of a function considering the combination of different respects such as relevance, coverage, coherence and diversity (Yan et al., 2011b). However, their approaches just treat timeline generation as a sentence ranking or optimization problem and seldom explore the topic information lied in the corpus.

Recently, topic models have been widely used for capturing the dynamics of topics via time. Many dynamic approaches based on LDA model (Blei et al., 2003) or Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006) have been proposed to discover the evolving patterns in the corpus as well as the snapshot clusters at each time epoch (Blei and Lafferty, 2006; Chakrabarti et al., 2006; Wang and McCallum, 2007; Caron et al., 2007; Ren et al., 2008; Ahmed and Xing, 2008; Zhang et al., 2010).

In this paper, we propose EHDP: a evolutionary hierarchical Dirichlet process (HDP) model for timeline summarization. In EHDP, each HDP is built for multiple corpora at each time epoch, and the time dependencies are incorporated into epochs under the Markovian assumptions. Topic popularity and topic-word distribution can be inferred from a Chinese Restaurant Process (CRP). Sentences are selected into timelines by considering different aspects such as topic relevance, coverage and coherence. We built the evaluation sys-
tems which contain 6 real datasets and performance of different models is evaluated according to the ROUGE metrics. Experimental results demonstrate the effectiveness of our model.

2 EHDP for Timeline Summarization

2.1 Problem Formulation

Given a general query \( Q = \{q_i\}_{i=1}^{Q_q} \), we firstly obtain a set of query related documents. We denote different corpus as \( C = \{C^t\}_{t=1}^{T} \) according to their published time where \( C^t = \{D_{i}^t\}_{i=1}^{N_t} \) denotes the document collection published at epoch \( t \). Document \( D_{i}^t \) is formulated as a collection of sentences \( \{s_{ij}^t\}_{j=1}^{N_{ti}} \). Each sentence is presented with a series of words \( s_{ij}^t = \{w_{ij}^t\}_{j=1}^{N_{ti}} \) and associated with a topic \( \theta_{ij}^t \). \( V \) denotes the vocabulary size. The output of the algorithm is a series of timelines summarization \( I = \{I_t\}_{t=1}^{T} \) where \( I^t \subset C^t \).

2.2 EHDP

Our EHDP model is illustrated in Figure 2. Specifically, each corpus \( C^t \) is modeled as a HDP. These HDP shares an identical base measure \( G_0 \), which serves as an overall bookkeeping of overall measures. We use \( G_0^t \) to denote the base measure at each epoch and draw the local measure \( G_t \) for each document at time \( t \) from \( G_0^t \). In EHDP, each sentence is assigned to an aspect \( \theta_{ij}^t \) with the consideration of words within current document.

To consider time dependency information in EHDP, we link all time specific base measures \( G_0^t \) with a temporal Dirichlet mixture model as follows:

\[
G_0^t \sim DP(\gamma^t, \frac{1}{K} G_0 + \frac{1}{K} \sum_{\delta=0}^{\Delta} F(v, \delta) \cdot G_{0}^{t-\delta})
\]

where \( F(v, \delta) = exp(-\delta/v) \) denotes the exponential kernel function that controls the influence of neighboring corpus. \( K \) denotes the normalization factor where \( K = 1 + \sum_{\delta=0}^{\Delta} F(v, \delta). \Delta \) is the time width and \( \lambda \) is the decay factor. In Chinese Restaurant Process (CRP), each document is referred to a restaurant and sentences are compared to customers. Customers in the restaurant sit around different tables and each table \( b_{in} \) is associated with a dish (topic) \( \Psi_{in} \) according to the dish menu. Let \( m_{tk} \) denote the number of tables enjoying dish \( k \) in all restaurants at epoch \( t \), \( m_{tk} = \sum_{i=1}^{N_t^k} \sum_{n=1}^{N_{ti}} 1(\Psi_{jn}^t = k) \). We redefine another parameter \( M_{tk} \) to incorporate time dependency into EHDP.

\[
M_{tk} = \sum_{\delta=0}^{\Delta} F(v, \delta) \cdot m_{t-\delta,k}
\]

Let \( n_{ib}^t \) denote the number of sentences sitting around table \( b \) in document \( i \) at epoch \( t \). In CRP for EHDP, when a new customer \( s_{ij}^t \) comes in, he can sit on the existing table with probability \( n_{ib}^t / (n_i^t - 1 + \gamma) \), sharing the dish (topic) \( \Psi_{ib}^t \) served at that table or picking a new table with probability \( \gamma / (n_i^t - 1 + \gamma) \). The customer has to select a dish from the global dish menu if he chooses a new table. A dish that has already been shared in the global menu would be chosen with probability \( M_{ik}^t / (\sum_k M_{ik}^t + \alpha) \) and a new dish with probability \( \alpha / (\sum_k M_{ik}^t + \alpha) \). For each epoch \( t \in [1, T] \):

1. draw global measure

\[
G_0^t \sim DP(\alpha, \frac{1}{K} G_0 + \frac{1}{K} \sum_{\delta=0}^{\Delta} F(v, \delta) \cdot G_{0}^{t-\delta})
\]

2. for each document \( D_{i}^t \) at epoch \( t \):

2.1 draw local measure \( G_{i}^t \sim DP(\gamma, G_{0}^t) \)

2.2 for each sentence \( s_{ij}^t \) in \( D_{i}^t \)

\[\text{draw aspect } \theta_{ij}^t \sim G_{i}^t \text{ for } w \in s_{ij}^t \text{ draw } w \sim f(w) \theta_{ij}^t \]

Figure 1: Generation process for EHDP
Figure 2: Graphical model of EHDP.

Relevance: the summary should be related with the proposed query Q.

$$F_R(I^t) = \zeta(KL(I^t||Q))$$

Coverage: the summary should highly generalize important topics mentioned in document collection at epoch t.

$$F_{C_v}(I^t) = \zeta(KL(I^t||C^t))$$

Coherence: News evolves over time and a good component summary is coherent with neighboring corpus so that a timeline tracks the gradual evolution trajectory for multiple correlated news.

$$F_{Ch}(I^t) = \sum_{\delta=-\Delta/2}^{\Delta/2} F(v, \delta) \cdot \zeta(KL(I^t||C^{t-\delta})) / \sum_{\delta=-\Delta/2}^{\Delta/2} F(v, \delta)$$

Let Score($I^t$) denote the score of the summary and it is calculated in Equ.(4).

$$Score(I^t) = \lambda_1 F_R(I^t) + \lambda_2 F_{C_v}(I^t) + \lambda_3 F_{Ch}(I^t)$$

$$\sum_1 \lambda_i = 1.$$ Sentences with higher score are selected into timeline. To avoid aspect redundancy, MMR strategy (Goldstein et al., 1999) is adopted in the process of sentence selection.

3 Experiments

3.1 Experiments set-up

We downloaded 3156 news articles from selected sources such as BBC, New York Times and CNN with various time spans and built the evaluation systems which contains 6 real datasets. The news belongs to different categories of Rule of Interpretation (ROI) (Kumaran and Allan, 2004). Detailed statistics are shown in Table 1. Dataset 2 (Deepwater Horizon oil spill), 3 (Haiti Earthquake) and 5 (Hurricane Sandy) are used as training data and the rest are used as test data. Summary at each epoch is truncated to the same length of 50 words.

Summaries produced by baseline systems and ours are automatically evaluated through ROUGE evaluation metrics (Lin and Hovy, 2003). For the space limit, we only report three ROUGE ROUGE-2-F and ROUGE-W-F score. Reference timeline in ROUGE evaluation is manually generated by using Amazon Mechanical Turk\(^1\). Workers were asked to generate reference timeline for news at each epoch in less than 50 words and we collect 790 timelines in total.

3.2 Parameter Tuning

To tune the parameters $\lambda(i = 1, 2, 3)$ and $v$ in our system, we adopt a gradient search strategy. We firstly fix $\lambda_1$ to 1/3. Then we perform experiments on with setting different values of $v/#epoch$ in the range from 0.02 to 0.2 at the interval of 0.02. We find that the Rouge score reaches its peak at round 0.1 and drops afterwards in the experiments. Next, we set the value of $v$ is set to 0.1 - $#epoch$ and gradually change the value of $\lambda_1$ from 0 to 1 with interval of 0.05, with simultaneously fixing $\lambda_2$ and $\lambda_3$ to the same value of $(1 - \lambda_1)/2$. The performance gets better as $\lambda_1$ increases from 0 to 0.25 and then declines. Then we set the value of $\lambda_1$ to 0.25 and change the value of $\lambda_2$ from 0 to 0.75 with interval of 0.05. And the value of $\lambda_2$ is set to 0.4, and $\lambda_3$ is set to 0.35 correspondingly.

3.3 Comparison with other topic models

In this subsection, we compare our model with 4 topic model baselines on the test data. Stand-HDP(1): A topic approach that models different time epochs as a series of independent HDPs without considering time dependency. Stand-HDP(2):

\(^1\)http://mturk.com

| News Source | Nation | News Source | Nation |
|-------------|--------|-------------|--------|
| BBC         | UK     | New York Times | US     |
| Guardian    | UK     | Washington Post | US     |
| CNN         | US     | Fox News     | US     |
| ABC         | US     | MSNBC       | US     |

Table 1: New sources of datasets

| News Subjects (Query) | #docs | #epoch |
|-----------------------|-------|--------|
| 1. Michael Jackson Death | 744   | 162    |
| 2. Deepwater Horizon oil spill | 642   | 127    |
| 3. Haiti Earthquake       | 247   | 83     |
| 4. American Presidential Election | 1246  | 286    |
| 5. Hurricane Sandy        | 317   | 58     |
| 6. Jerry Sandusky Sexual Abuse | 320   | 74     |

Table 2: Detailed information for datasets
A global HDP which models the whole time span as a restaurant. The third baseline, Dynamic-LDA is based on Blei and Lafferty’s work and Stan-LDA is based on standard LDA model. In LDA based models, aspect number is predefined as 80. Experimental results of different models are shown in Table 2. As we can see, EHDP achieves better results than the two standard HDP baselines where time information is not adequately considered. We also find an interesting result that Stan-HDP performs better than Stan-LDA. This is partly because new aspects can be automatically detected in HDP. As we know, how to determine topic number in the LDA-based model is still an open problem.

### 3.4 Comparison with other baselines

We implement several baselines used in traditional summarization or timeline summarization for comparison. (1) 

**Centroid** applies the MEAD algorithm (Radev et al., 2004) according to the features including centroid value, position and first-sentence overlap. (2) **Manifold** is a graph based unsupervised method for summarization, and the score of each sentence is got from the propagation through the graph (Wan et al., 2007). (3) **ETS** is the timeline summarization approach developed by Yan et al., (2011a), which is a graph based approach with optimized global and local biased summarization. (4) **Chieu** is the timeline system provided by (Chieu and Lee, 2004) utilizing interest and bursty ranking but neglecting trans-temporal news evolution. As we can see from Table 3, Centroid and Manifold get the worst results. This is probably because methods in multi-document summarization only care about sentence selection and neglect the novelty detection task. We can also see that EHDP under our proposed framework outputs existing timeline summarization approaches ETS and chieu. Our approach outputs Yan et al.,(2011a)s model by 6.9% and 9.3% respectively with regard to the average score of ROUGE-2-F and ROUGE-W-F.

### 4 Conclusion

In this paper we present an evolutionary HDP model for timeline summarization. Our EHDP extends original HDP by incorporating time dependencies and background information. We also develop an effective sentence selection strategy for candidate in the summaries. Experimental results on real multi-time news demonstrate the effectiveness of our topic model.

### References

Amr Ahmed and Eric Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process. 2008. In *SDM*.
James Allan, Rahul Gupta and Vikas Khandelwal. Temporal summaries of new topics. 2001. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.

David Blei, Andrew Ng and Michele Jordan. 2003. Latent dirichlet allocation. In Journal of Machine Learning Research.

David Blei and John Lafferty. Dynamic topic models. 2006. In Proceedings of the 23rd international conference on Machine learning.

François Carol, Manuel Davy and Arnaud Doucet. Generalized poly urn for time-varying dirichlet process mixtures. 2007. In Proceedings of the International Conference on Uncertainty in Artificial Intelligence.

Deepayan Chakrabarti, Ravi Kumar and Andrew Tomkins. Evolutionary Clustering. In Proceedings of the 12th ACM SIGKDD international conference Knowledge discovery and data mining.

Hai-Leong Chieu and Yoong-Keok Lee. Query based event extraction along a timeline. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In Proceedings of the 27th annual international ACM SIGIR04.

Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha and Yong Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In Proceedings of the 18th international conference on World wide web.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the Human Language Technology Conference of the NAACL. 2003.

Dragomir Radev, Hongyan. Jing, and Malgorzata Stys. 2004. Centroid-based summarization of multiple documents. In Information Processing and Management.

Lu Ren, David Dunson and Lawrence Carin. The dynamic hierarchical Dirichlet process. 2008. In Proceedings of the 25th international conference on Machine Learning.

Xiaojun Wan, Jianwu Yang and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In Proceedings of International Joint Conference on Artificial Intelligence.

Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.