Unsupervised Storyline Extraction from News Articles

Deyu Zhou[†‡] **Haiyang Xu**[†] **Xin-Yu Dai**[‡] **Yulan He**[§]
[†] School of Computer Science and Engineering, Southeast University, China
[‡] State Key Laboratory for Novel Software Technology, Nanjing University, China
[§] School of Engineering and Applied Science, Aston University, UK
{d.zhou,h.xu}@seu.edu.cn, daixinyu@nju.edu.cn, y.he@cantab.net

Abstract

Storyline extraction from news streams aims to extract events under a certain news topic and reveal how those events evolve over time. It requires algorithms capable of accurately extracting events from news articles published in different time periods and linking these extracted events into coherent stories. The two tasks are often solved separately, which might suffer from the problem of error propagation. Existing unified approaches often consider events as topics, ignoring their structured representations. In this paper, we propose a non-parametric generative model to extract structured representations and evolution patterns of storylines simultaneously. In the model, each storyline is modelled as a joint distribution over some locations, organizations, persons, keywords and a set of topics. We further combine this model with the Chinese restaurant process so that the number of storylines can be determined automatically without human intervention. Moreover, per-token Metropolis-Hastings sampler based on light latent Dirichlet allocation is employed to reduce sampling complexity. The proposed model has been evaluated on three news corpora and the experimental results show that it outperforms several baseline approaches.

1 Introduction

With the rapid development of online news media sites, tremendous news reports are generated each day. Moreover, newsworthy events are widely scattered not only on traditional news media but also on social media. It is crucial to develop an automated tool which can provide a temporal summary of related events and their evolutions from web texts. Therefore, storyline extraction, aiming at summarising the development of certain related events, has been extensively studied in recent years.

Considering each storyline as a cluster, storyline extraction can be treated as an evolutionary clustering problem, where documents are ordered as a stream and clustered depending on both similarity in context and closeness in time. Several methods have been proposed, differing in the ways of calculating context similarity and time closeness [Yan *et al.*, 2011; Huang and Huang, 2013; Du *et al.*, 2015]. However, most of the previous approaches do not represent storylines in the form of structured representations. Moreover, in the approaches based on Bayesian nonparametric models [Li and Cardie, 2014; Diao and Jiang, 2014], events are extracted independently from different time periods and are subsequently combined as storylines in a post-processing step. On the contrary, we propose a model which is able to capture the dynamic aspect of storylines from text stream data. In specific, we model each storyline as a joint distribution over some locations l, organizations o, persons p, keywords w and a set of topics z. Each word in a document is discriminated as storyline keyword, background word and global topic word to learn storylines' representations more accurately. Moreover, storyline-keyword, storyline-location, storyline-person and storyline-organization probabilities at epoch t are dependent on their corresponding probabilities in the last M epochs in order to capture the dynamic evolution of different events in the same storyline over time. We further combine this model with the Chinese restaurant process so that the number of storylines in each epoch can be determined automatically without human intervention. Moreover, per-token Metropolis-Hastings sampler based on lightLDA [Yuan *et al.*, 2015] is employed to reduce sampling complexity.

The main contributions of the paper are summarized below:

- We propose a novel non-parametric generative model to extract structured representations and evolution patterns of storylines simultaneously. The model is combined with the Chinese restaurant process to automatically determine the number of storylines in each epoch without human intervention.
- Per-token Metropolis-Hastings sampler based on lightLDA is employed to reduce sampling complexity.
- The proposed approach has been evaluated on three datasets and a significant improvement on F-measure compared to the state-of-the-art approaches is observed.

2 Related Work

There have been many studies on storyline extraction from text including news articles. For example, to generate sto-

ryline from massive texts on the Internet, [Yan *et al.*, 2011] calculated correlations of individual summaries on each date generated using a text summarization algorithm. [Kawamae, 2011] proposed a trend analysis model by using the difference between temporal words and other words in each document to detect topic evolution over time. [Lin *et al.*, 2012] extracted a storyline by first obtaining relevant tweets and then generating storylines via graph optimization. In [Binh Tran *et al.*, 2013], relevant time points and contents to be included in the storyline summary were chosen based on a linear regression model. [Radinsky and Horvitz, 2013] constructed storylines based on text clustering and entity entropy. [Huang and Huang, 2013] developed a mixture event aspect model to distinguish local and global aspects of events described in sentences and utilized an optimization method to generate storylines. [Huang *et al.*, 2015] first extracted topics from a short text corpus based on word co-occurrence patterns, and then developed an event evolution mining algorithm to discover hot events and their evolutions. [Vossen *et al.*, 2015] proposed a model based on the narratology framework by constructing storylines from events through bridging relations. [Yu *et al.*, 2015] proposed a context-dependent news article storyline detection based on dense subgraph learning, which adaptively learns a set of cross-article link patterns.

With the popularity of using topic models for document clustering [Blei *et al.*, 2003], Bayesian nonparametric models, such as Dirichlet process (DP) [Li and Cardie, 2014] and recurrent Chinese Restaurant process (RCRP) [Diao and Jiang, 2014], attract many research attentions for tackling the storyline extraction task. [Li and Cardie, 2014] proposed a non-parametric multi-level DP model to reconstruct users' life history based on their Twitter streams. [Wang, 2013] proposed a time-dependent hierarchical DP for storyline generation. It can detect different levels of topic information across corpus and the structure is further used for sentence selection. [Zhang *et al.*, 2010] proposed an evolutionary hierarchical DPs to discover interesting cluster evolution patterns from texts by adding time dependencies to the adjacent epochs. In [Li and Li, 2013], an evolutionary hierarchical DP was proposed to capture the topic evolution pattern in storyline summarization. Instead of DP, [Ahmed *et al.*, 2011] proposed a dynamic nonparametric model based on RCRP to group temporally and topically related news articles into same storylines in order to reveal the temporal evolution of events. Following this way, [Tang and Yang, 2012] developed a topic user trend model, which incorporates user interests into the generative process of web contents. [Diao and Jiang, 2014] proposed a probabilistic model to identify both events and topics simultaneously from Twitter by applying a duration-based probability discount to RCRP to capture the character of events on Twitter. [Du *et al.*, 2015] combined Dirichlet processes and Hawkes processes to deal with asynchronous streaming data in an online manner. [Blei and Frazier, 2011] developed the distance dependent Chinese Restaurant process (CRP) to model dependencies between data in infinite clustering models including dependencies across time or space. Following this way, [Tang *et al.*, 2015] proposed a hybrid distance dependent CRP based hierarchical topic model and utilized it as prior for news article clustering.

[Zhou *et al.*, 2015] proposed an unsupervised Bayesian model for storyline extraction. They assumed each document belongs to one storyline s, which is modelled as a joint distribution over some named entities e and a set of topics z. The weighted sum of the storyline distribution of previous epochs is used to set the prior of the storyline distribution in the current epoch. However, the model requires the number of storylines to be preset. In reality, it is impossible to know this information in advance. Furthermore, their model did not distinguish between the global topic words shared by the whole corpus and storyline-specific keywords.

3 Methodology

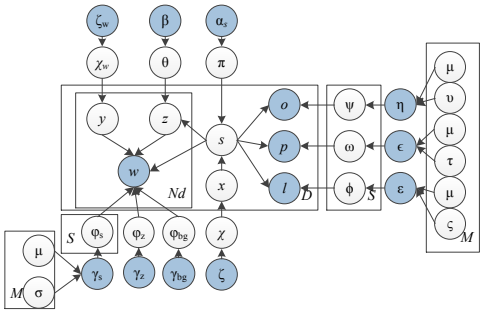

Figure 1: Dynamic Storyline Extraction model (DSEM).

To model the generation of a storyline in consecutive time periods from a stream of documents, we propose an unsupervised latent variable model, called dynamic storyline extraction Model (DSEM). The graphical model of DSEM is shown in Figure 1. The model has such features: (1) for a certain period of time, we assume that each document could belong to one storyline s and each storyline s is modelled as a joint distribution over some locations l, organizations o, persons p, keywords w and a set of topics z; (2) the story described in the document d at current time t is drawn either from an existing storyline or a new storyline at current time t. The latent random variable x_d^t acts as a storyline-category switch; (3) to represent the dependency of different stories in the same storyline, storyline-word, storyline-location, storyline-person, storyline-organization probabilities at time t are dependent on the previous corresponding distributions in the last M epochs; (4) each word $w_{d,n}$ in the document d is drawn either from the storyline key word distribution, or the background word distribution, or the global-topic word distribution shared by the corpus. A switch variable $y_{d,n}$ is used to control word generation process; (5) the number of storylines is determined automatically by applying CRP at current t. CRP is a non-parametric model which can be used for clustering. To model streaming data, CRP models a restaurant with infinite number of tables and each with infinite capacity. When the $i-th$ customer comes in, she can choose an already occupied table with probability $n_k^{(i)}/(n-1+\alpha)$ or a new table with probability $\alpha/(n-1+\alpha)$, where $n_k^{(i)}$ is

the number of customers sitting at table k and n is the total number of customers in this restaurant.

The generative process of DSEM is shown as below. For each time period t from 1 to T:

- Draw a distribution over storyline switch $\chi^t \sim$ Beta(ζ^t), draw a distribution over word switch $\chi^t_w \sim$ Dirichlet(ζ^t_w).
- Draw a distribution over prior storylines $\pi^t_s \sim$ Dirichlet(α^t_s), draw a distribution over global topics $\theta^t \sim$ Dirichlet(β^t).
- For each topic $k \in \{1...K\}$, draw a word distribution $\varphi^t_z \sim$ Dirichlet(γ^t_z), For background words, draw a word distribution $\varphi^t_{bg} \sim$ Dirichlet(γ^t_{bg}).
- For each existing storyline $s \in \{1..S_t\}$:
  - Draw a distribution over location $\phi^t_s \sim$ Dirichlet(ε^t_s), person $\omega^t_s \sim$ Dirichlet(ϵ^t_s), organization $\psi^t_s \sim$ Dirichlet(η^t_s),
  - For storyline keyword, draw a word distribution $\varphi^t_s \sim$ Dirichlet(γ^t_s).
- For each document $d \in \{1..D_t\}$:
  - Choose $x^t_d \sim$ Multinomial(χ^t),
  - If $x^t_d = 0$, choose a new storyline indicator s^t_d from CRP; draw Multinomial distribution $\phi^0_s, \omega^0_s, \varphi^0_s$ respectively for this new storyline,
  - If $x^t_d = 1$, choose a storyline indicator s^t_d from Multinomial(π^t_s),
  - For each location $l \in \{1..L_d\}$, Choose a location $l \sim$ Multinomial(ϕ^t_s),
  - For each person $p \in \{1..P_d\}$, Choose a person $p \sim$ Multinomial(ω^t_s),
  - For each organization $o \in \{1..O_d\}$, Choose an organization $o \sim$ Multinomial(ψ^t_s),
  - For other word positions $n \in \{1..N_d\}$:
    - Choose $y_{d,n} \sim$ Multinomial(χ^t_w),
    - If $y_{d,n} = 0$, choose a background word $w_{n,bg} \sim$ Multinomial(φ^t_{bg}),
    - If $y_{d,n} = 1$, choose a storyline keyword $w_{n,s} \sim$ Multinomial(φ^t_s),
    - If $y_{d,n} = 2$, choose a global topic $z_n \sim$ Multinomial(θ^t), choose a word $w_{n,z} \sim$ Multinomial(φ^t_z).

To represent the dependency of stories belonging to the same storyline, we define a vector of $M+1$ weights $\mu^t_s = \{\mu^t_{s,m}\}^M_{m=0} (\mu^t_{s,m} > 0)$ that each $\sigma^t_{s,m}, \varsigma^t_{s,m}, \tau^t_{s,m}, \upsilon^t_{s,m}$ contributes to calculating the priors of $\varphi^t_s, \phi^t_s, \omega^t_s, \psi^t_s$, where $\sigma^t_{s,m}$ denotes the storyline-keyword distribution under storyline s at previous epoch m, $\varsigma^t_{s,m}$ denotes the storyline-location distribution under storyline s at previous epoch m, $\tau^t_{s,m}$ denotes the storyline-person distribution under storyline s at previous epoch m and $\upsilon^t_{s,m}$ denotes the storyline-organization distribution under storyline s at previous epoch m.

The priors for storyline-keyword distribution φ^t_s, storyline-location distribution ϕ^t_s, storyline-person distribution ω^t_s, storyline-organization distribution ψ^t_s at epoch t are calculated as $\gamma^t_s = \sum^M_{m=0} \mu^t_{s,m}\sigma^t_{s,m}, \varepsilon^t_s = \sum^M_{m=0} \mu^t_{s,m}\varsigma^t_{s,m}, \epsilon^t_s = \sum^M_{m=0} \mu^t_{s,m}\tau^t_{s,m}, \eta^t_s = \sum^M_{m=0} \mu^t_{s,m}\upsilon^t_{s,m}$.

In our experiments, the weight parameters are set to be the same regardless of storylines. They are only dependent on the time window using an exponential decay function, $\mu_m = exp(-\lambda \cdot m)$ where m stands for the m_{th} epoch counting backwards in the past M epoches. That is, more recent documents would have a relatively stronger influence on the model parameters in the current epoch compared to earlier documents.

The storyline-keyword distribution φ^t_s, the storyline-location distribution ϕ^t_s , the storyline-person distribution ω^t_s and the storyline-organization distribution ψ^t_s at epoch t are generated from the Dirichlet distribution parameterized by $\gamma^t_s, \varepsilon^t_s, \epsilon^t_s, \eta^t_s$, $\varphi^t_s \sim Dir(\gamma^t_s), \phi^t_s \sim Dir(\varepsilon^t_s), \omega^t_s \sim Dir(\epsilon^t_s)$, and $\psi^t_s \sim Dir(\eta^t_s)$, respectively. With this formulation, we can ensure that the mean of the Dirichlet parameter for the current epoch is proportional to the weighted sum of the storyline-keyword, storyline-location, storyline-person, storyline-organization distributions at previous epochs.

4 Inference and Parameter Estimation

We use collapsed Gibbs sampling to infer the parameters of the model, given observed data. Gibbs sampling is a Markov chain Monte Carlo method which allows us to repeatedly sample from a Markov chain whose stationary distribution is the posterior of interest, s^t_d and $z^t_{d,n}$ here, from the distribution over that variable given the current values of all other variables and the data. Such samples can be used to empirically estimate the target distribution.

4.1 Storyline Sampling

Letting the subscript $-d$ denote the quantity that excludes counts in document d, the conditional posterior for s^t_d is:

$$p(s^t_d = j|\overrightarrow{s^t_{-d}}, \overrightarrow{z}, \overrightarrow{w}, \overrightarrow{p}, \overrightarrow{l}, \Lambda) \propto$$

$$\frac{\prod^L_l \prod^{n^{(d)}_{j,l}}_{b=1}(n_{j,l} + \varepsilon^t_{j,l} - b)}{\prod^{L_{(d)}}_{b=1}(n^L_j + \sum^L_{l=1}\varepsilon^t_{j,l} - b)} \cdot \frac{\prod^O_o \prod^{n^{(d)}_{j,o}}_{b=1}(n_{j,o} + \eta^t_{j,o} - b)}{\prod^{O_{(d)}}_{b=1}(n^O_j + \sum^O_{o=1}\eta^t_{j,o} - b)}$$

$$\cdot \frac{\prod^P_p \prod^{n^{(d)}_{j,p}}_{b=1}(n_{j,p} + \epsilon^t_{j,p} - b)}{\prod^{P_{(d)}}_{b=1}(n^P_j + \sum^P_{p=1}\epsilon^t_{j,p} - b)} \cdot \frac{\prod^V_v \prod^{n^{(d)}_{j,v}}_{b=1}(n_{j,v} + \gamma^t_{j,v} - b)}{\prod^{V_{(d)}}_{b=1}(n^V_j + \sum^V_{v=1}\gamma^t_{j,v} - b)}$$

$$\cdot \frac{\prod^K_k \prod^{n^{(d)}_{j,k}}_{b=1}(n_{j,k} + \beta^t_k - b)}{\prod^{K_{(d)}}_{b=1}(n^K_j + \sum^K_{k=1}\beta^t_k - b)} \cdot \frac{\prod^V_v \prod^{n^{(d)}_{k,v}}_{b=1}(n_{k,v} + \gamma^t_{k,v} - b)}{\prod^{V_{(d)}}_{b=1}(n^V_k + \sum^V_{v=1}\gamma^t_{k,v} - b)}$$

$$\cdot \begin{cases} \frac{n^t_{new,-d} + \zeta^t_{new}}{D_t + \zeta^t_{new} + \zeta^t_{pri}} \cdot \frac{\alpha}{n^t_{new,-d} + \alpha}, & \text{new storyline at } t \\ \frac{n^t_{new,-d} + \zeta^t_{new}}{D_t + \zeta^t_{new} + \zeta^t_{pri}} \cdot \frac{n^t_{new,j,-d}}{n^t_{new,-d} + \alpha}, & \text{existing storyline at } t \\ \frac{n^t_{pri,-d} + \zeta^t_{pri}}{D_t + \zeta^t_{new} + \zeta^t_{pri}} \cdot \frac{n^t_{pri,j,-d} + \alpha_j}{\sum^{S_t}_{s=1}(n_s + \alpha_s - 1)}, & \text{other} \end{cases}$$

where D_t is the total number of documents in current epoch t, n^t_{new} denotes the number of documents assigned to the new storyline generated in current epoch t, $n^t_{new,j}$ denotes the number of documents assigned to new storyline indicator j in current epoch t, n^t_{pri} denotes the number of documents assigned to an existing storyline generated in the past M epochs, $n^t_{pri,j}$ denotes the number of documents assigned to an existing storyline indicator j generated in the past M epochs, $n_{j,l}$ is the number of times location l assigned with storyline j, n^L_j denotes the total number of locations with storyline j in the document collection, $n_{j,o}$ is the number of times organization o assigned with storyline j, n^O_j denotes the total number of organizations with storyline j in the document collection, $n_{j,p}$ is the number of times person p assigned with storyline j, n^P_j denotes the total number of persons with storyline j in the document collection, $n_{j,k}$ is the number of times words with topic label k with storyline j, n^K_j denotes the total number of words with global topic with storyline j in the document collection, $n_{k,v}$ is the number of times the word v with topic k, n^V_k denotes the total number of words with global topic in the document collection and counts with (d) notation denote the counts relating to the document d only. The terms in the big curly bracket denote the probability of assigning document d to storyline j by incorporating CRP.

4.2 Word Switch Variable and Topic Sampling

For each word token $w_{d,n}$ in document d, the posterior probability of adding it to background word is derived as:

$$p(y_{d,n} = 0|\overrightarrow{y_{d,-n}}, \overrightarrow{w_{bg}}, \Lambda) \propto \frac{n_{m,bg} + \zeta^t_{bg} - 1}{\sum^3_{y=1}(n_{m,y} + \zeta^t_y) - 1} \cdot \frac{n_{bg,w_{d,n}} + \gamma^t_{bg,w_{d,n}} - 1}{\sum^V_{v=1}(n_{bg,v} + \gamma^t_{bg,v}) - 1}$$

where $n_{m,bg}$ denotes the number of words in document m assigned to the background word and $n_{bg,v}$ denotes the number of word v assigned to the background word.

For each word token $w_{d,n}$ in document d, the posterior probability of adding it to storyline j key word is derived as:

$$p(y_{d,n} = 1|\overrightarrow{y_{d,-n}}, \overrightarrow{w_j}, \Lambda) \propto \frac{n_{m,j} + \zeta^t_s - 1}{\sum^3_{y=1}(n_{m,y} + \zeta^t_y) - 1} \cdot \frac{n_{j,w_{d,n}} + \gamma^t_{j,w_{d,n}} - 1}{\sum^V_{v=1}(n_{j,v} + \gamma^t_{j,v}) - 1}$$

where $n_{m,j}$ denotes the number of words in document m assigned to the storyline j key word and $n_{j,v}$ denotes the number of word v assigned to storyline j key word.

For each word token $w_{d,n}$ in document d , the posterior probability of adding word to global topic k with storyline j is derived as:

$$p(y_{d,n} = 2, z^t_{d,n} = k|\overrightarrow{y_{d,-n}}, \overrightarrow{w_{j,z}}, \Lambda) \propto \frac{n_{m,z} + \zeta^t_z - 1}{\sum^3_{y=1}(n_{m,y} + \zeta^t_y) - 1} \cdot \frac{n_{k,w_{d,n}} + \gamma^t_{k,w_{d,n}} - 1}{\sum^V_{v=1}(n_{k,v} + \gamma^t_{k,v}) - 1} \cdot \frac{n_{j,k} + \beta^t_k - 1}{\sum^K_{c=1}(n_{j,c} + \beta^t_c) - 1}$$

where $n_{m,z}$ denotes the number of words in document m assigned to the global topic, $n_{k,v}$ denotes the number of word v assigned to the global topic k and $n_{j,c}$ denotes the number of words assigned to the global topic c with storyline j.

The hyperparameters of the model are set $\alpha = 1, \lambda = 0.5, \alpha^t_s = \beta^t = \zeta^t_{bg} = 0.1, \zeta^t_{pri} = \zeta^t_{new} = \gamma^t_z = \gamma^t_{bg} = 0.01, \zeta^t_s = 0.2, \zeta^t_z = 0.7 (s \in 1..S_t, t \in 1..T)$ in our experiment.

5 Reduction of Sampling Complexity

In our model, the Gibbs sampling procedure mainly consists of two steps. The first step is storyline sampling, which scales linearly with the number of words in the document requiring a computation for every existing storyline. To solve this problem, we turn to Metropolis Hastings (MH) and define $q(s)$ as the storyline proposal distribution

$$q(s) = p(s^t_d|\Lambda)p(l^t_d|s^t_d, \Lambda)p(p^t_d|s^t_d, \Lambda)p(o^t_d|s^t_d, \Lambda)$$

whose computation complexity scales linearly with the number of important entities such as location, person and organization. We sample s^* from this proposal and compute the acceptance rate r of state $s^t_d \to s^*$.

The second step is topic sampling, which needs $O(K)$ per-token complexity. Here, we borrow the idea of LightLDA [Yuan *et al.*, 2015] to reduce sampling complexity, in which a new $O(1)$ per-token MH sampler has been developed. We employ two proposal distributions and combine both proposals to improve mixing. The first proposal is word-proposal:

$$q_w(k) \propto \frac{n_{k,v} + \gamma^t_{k,v} - 1}{\sum^V_{v=1}(n_{k,v} + \gamma^t_{k,v}) - 1}$$

where $n_{k,v}$ denotes the number of word v assigned to the global topic k. Once $z \sim q_w(z)$ is sampled, we can compute the acceptance probability in $O(1)$ time, as long as we keep track of all sufficient statistics n during sampling. To sample from q_w in $O(1)$, We use the alias approach, which transforms a non-uniform distribution into a uniform distribution. Furthermore, The building cost of alias table gets amortized to $O(1)$ due to alias table re-used in MH sampling.

The second proposal is storyline-proposal:

$$q_s(k) \propto n_{j,k} + \beta^t_k$$

where $n_{j,k}$ denotes the number of words assigned to the global topic k with storyline j. We can sample from $n_{j,k}$ by simply drawing an integer j uniformly from $1, 2, ..., n_j$, so the storyline-proposal can be sampled in $O(1)$ non-amortized time and we do not need to construct an alias table.

6 Experiments

In this section, we first introduce the datasets we used for our experiments, and then describe the baseline approaches, finally present the experimental results.

6.1 Datasets

We crawled and parsed the GDELT Event Database (http://data.gdeltproject.org/events/index.html) containing news articles published in the month of May in 2014. The

whole one-month data contain 526,587 documents, which we called Dataset I. As it is time consuming to identify all the true storylines in such a large dataset, we only report the precision but not recall of the storylines extracted. To fully evaluate the performance of the proposed approach, we manually annotated one-week data extracted from Dataset I, containing 101,654 documents, which we called Dataset II. In dataset II, altogether 77 storylines are identified. In general, storylines can be categorized into four types: (1) long-term storylines which last more than 2 weeks; (2) short-term storylines which last less than 1 week; (3) intermittent storylines which last more than 2 weeks and interrupt in the middle period; (4) new storylines which start in the middle of the period, not the beginning. We found that in Dataset II, not all these types of storylines exist. As the proposed approach aims to identify all types of storylines, we therefore manually construct Dataset III by random selecting all types of storylines from Dataset I. This dataset consists 23,376 documents annotated with 30 storylines.

In our experiments, we used the Stanford Named Entity Recognizer for identifying the named entities. In addition, we removed common stopwords and only kept tokens which are verbs, nouns, or adjectives from these news documents. For the purpose of modelling, we split news documents by their publication dates and consider each day as an epoch.

6.2 Baselines

We chose the following methods as the baselines.

1. K_Means+Cosine_Similarity (KMCS): It first applies K_Means to cluster documents in each epoch, then uses Cosine_Similarity to link stories in different epochs to form a storyline.
2. Stan-LDA+Cosine_Similarity (LDCS): It first applies standard latent Dirichlet allocation (LDA) to detect latent storyline indicator. Each storyline is modelled as a joint distribution over some named entities and words. Cosine_Similarity is used to link stories in different epochs to form a storyline.
3. Dyn-LDA (DLDA): Topic-word distribution and popularity are linked across epochs by make a Markovian assumption in the model.
4. RCRP [Ahmed *et al.*, 2011]: It is a non-parameter model for evolutionary clustering, which assumes that the past story popularity is a good prior for current popularity.
5. SDM [Zhou *et al.*, 2015]: It assumes that the number of storylines is a constant and the storyline is modelled as a joint distribution over entities and keywords. The dependency of different stories of the same storyline at different epochs is captured by the modification of Dirichlet priors.

For SDM, the storyline number is set to 100 on Dataset II and 30 on Dataset III. The topic number is set to 100 on Dataset II and 20 on Dataset III. The number of historical epochs M, which is taken into account for setting the Dirichlet priors for the storyline-keyword, storyline-location, storyline-person, storyline-organization distributions, is set to 7, the same as in our proposed approach. For RCPR, the hyperparamter α is set to 1. Both LDCS and DLDA employ the same storyline number and topic number as SDM. For KMCS, the number of clusters is set to 100 on Dataset II and 30 on Dataset III.

6.3 Evaluation Metrics

To evaluate the performance of the proposed approach, we use precision, recall and F-measure which are commonly used in evaluating information extraction systems. The precision is calculated based on the following criteria: 1) The entities and keywords extracted refer to the same storyline. 2) The duration of the storyline is correct. We assume that the start date (or end date) of a storyline is the publication date of the first (or last) news article about it.

6.4 Experimental Results

The evaluation results of our proposed approach in comparison to the baselines on dataset I, II and III are presented in Table 1. For Dataset I, as it is hard to know the ground-truth of storylines within it, we only report the precision value by manually examining the extracted storylines. It can be observed from Table 1 that simply using K-means to cluster news articles in each day and linking similar stories across different days in hoping of identifying storylines gives the worst results on both dataset II and III. Using LDA (LDCS) to detect stories in each day improves the precision dramatically. The dynamic LDA model (DLDA) assumes topics (or stories) in the current epoch evolve from the previous epoch and further improves the storyline detection results significantly. SDM aims to capture the long distance dependencies in which the statistics gathered in the past M days are taken into account to set the Dirichlet priors of the storyline-topic-word, storyline-topic and storyline-entity distributions in the current epoch. However, SDM cannot generate new storylines. While RCRP can generate new storylines automatically, it does not model explicitly the generation of named entities such as persons, locations and organizations. As a result, it failed to identify any intermittent storylines. We also notice that SDM gives better results compared to RCRP on Dataset II but performs worse on Dataset III. One possible reason is the Dataset III contains some storylines happened sometime in the middle of the time period processed. SDM expects all the storylines start around the same time and hence is not able to detect new storylines happened at a later time.

By considering not only the long distance dependencies of storylines but also the new storylines generated in the middle of the time period, our proposed approach gives the best precision value on Dataset I and achieves the best performance on both dataset II and III. In specific, for Dataset II containing 77 storylines, our proposed model extracted 82 storylines among which 60 are correct. For Dataset III consisting of 30 storylines, our proposed model extracted 28 storylines among which 21 are correct. Moreover, the number of correctly extracted storylines by the proposed approach is more than that extracted by SDM. It empirically verifies that the proposed model can handle different types of storylines more effectively compared with the baselines when dealing with big data stream.

Table 1: Performance comparison of the storyline extraction results on Dataset I, II and III.

Dataset I		
Method	Precision(%)	# of extracted storylines
SDM	70.2	104
Our model	**75.43**	**114**
Dataset II		
Method	Precision(%)	Recall(%)	F-measure(%)
KMCS	22.73	32.47	26.74
LDCS	34.29	31.17	32.66
DLDA	62.67	61.03	61.84
RCRP	67.11	66.23	66.67
SDM	70.67	68.80	69.27
Our model	**73.17**	**77.92**	**75.47**
Dataset III		
Method	Precision(%)	Recall(%)	F-measure(%)
KMCS	20	16.67	18.57
LDCS	23.08	20	21.43
DLDA	46.16	43.33	42.86
RCRP	61.54	53.33	57.14
SDM	54.17	43.33	48.15
Our model	**75**	**70**	**72.41**

6.5 Structured Browsing

We illustrate the evolution of storylines by using structured browsing, from which the structured information (entities, topics, keywords) about storylines, the duration of storylines and the storyline popularity over time can be easily observed. Figure 2 shows the extracted storyline related to "India Election". Six representative epochs are highlighted in Figure 2 with detailed structured representations of the storyline. At the beginning, it can be easily deduced that there will be a prime minister election between Modi and Gandhi in India from the structured representation p "Modi, Gandhi" and w "prime, elect" at day 1. After that, election commission denied Modi's rally at day 9 which can be found from p "Modi", o "commission" and w "deny, rally". Then, at day 13, exit polls predicted that Modi will be next prime minister of India. At day 16, Modi won election to be prime minister of India and attended his swearing ceremony at day 21. Finally, Modi set up the new cabinet at day 26. All these information can be easily found from the structured representations of the storyline generated by our proposed approach. Based on the topic related words shown in Figure 2, we can also easily classify the extracted storyline as "politics".

6.6 Computation Complexity

To explore the efficiency of the Metropolis-Hastings sampler for the reduction of sampling complexity of the proposed model, we conducted an experiment by running the proposed approach with and without using the Metropolis-Hastings method. We run experiments on the dataset which consists of 500 documents initially. After that, the dataset is enlarged by gradually adding 500 documents each time until it contains 10,000 documents. The topic number is set to 50. Figure 3 illustrates the time consumed in each iteration when trained on the datasets with different sizes. It can be observed that the running time of both approaches increases

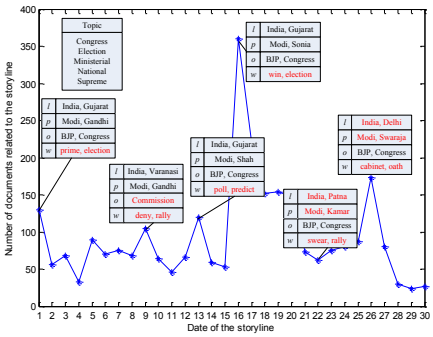

Figure 2: The structured representations of the storyline "India Election".

with the increasing size of the data. But the running time of the proposed approach without Metropolis-Hastings sampler is about 6 times slower compared to the approach using it.

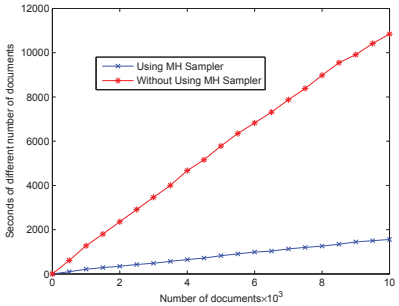

Figure 3: Comparison of training time consumed with or without using the Metropolis-Hastings sampler.

7 Conclusions and Future Work

In this paper, we have proposed an unsupervised Bayesian model to extract storylines from news corpus. We further combine this model with CRP so that the number of storylines in each epoch can be determined automatically without human intervention. Moreover, per-token MH sampler based on lightLDA is employed to reduce sampling complexity. Experimental results show that our proposed model outperforms a number of baselines. In future work, we will explore the impact of different scales of the dependencies from historical epochs on the performance of storyline extraction.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (61528302, 61472183), the Innovate UK under the grant number 101779 and the Collaborative Innovation Center of Wireless Communications Technology.

References

[Ahmed *et al.*, 2011] Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J Smola, and Choon Hui Teo. Unified analysis of streaming news. In *Proceedings of the 20th international conference on World wide web*, pages 267–276. ACM, 2011.

[Binh Tran *et al.*, 2013] Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. Predicting relevant news events for timeline summaries. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 91–92, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[Blei and Frazier, 2011] David M. Blei and Peter I. Frazier. Distance dependent chinese restaurant processes. *J. Mach. Learn. Res.*, 12:2461–2488, November 2011.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[Diao and Jiang, 2014] Qiming Diao and Jing Jiang. Recurrent chinese restaurant process with a duration-based discount for event identification from twitter. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 388–397, 2014.

[Du *et al.*, 2015] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J. Smola, and Le Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 219–228, 2015.

[Huang and Huang, 2013] Lifu Huang and Lian'en Huang. Optimized event storyline generation based on mixture-event-aspect model. In *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*, pages 726–735. ACL, 2013.

[Huang *et al.*, 2015] Guangyan Huang, Jing He, Yanchun Zhang, Wanlei Zhou, Hai Liu, Peng Zhang, Zhiming Ding, Yue You, and Jian Cao. Mining streams of short text for analysis of world-wide event evolutions. *World Wide Web*, 18:1201–1217, 2015.

[Kawamae, 2011] Noriaki Kawamae. Trend analysis model: trend consists of temporal words, topics, and timestamps. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 317–326. ACM, 2011.

[Li and Cardie, 2014] Jiwei Li and Claire Cardie. Timeline generation: Tracking individuals on twitter. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 643–652, New York, NY, USA, 2014. ACM.

[Li and Li, 2013] Jiwei Li and Sujian Li. Evolutionary hierarchical dirichlet process for timeline summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 556–560. ACL, 2013.

[Lin *et al.*, 2012] Chen Lin, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, and Tao Li. Generating event storylines from microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 175–184. ACM, 2012.

[Radinsky and Horvitz, 2013] Kira Radinsky and Eric Horvitz. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM, 2013.

[Tang and Yang, 2012] Xuning Tang and Christopher C Yang. TUT: a statistical model for detecting trends, topics and user interests in social media. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 972–981. ACM, 2012.

[Tang *et al.*, 2015] Siliang Tang, Fei Wu, Si Li, Weiming Lu, Zhongfei Zhang, and Yueting Zhuang. Sketch the storyline with charcoal: A non-parametric approach. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 3841–3848. AAAI Press, 2015.

[Vossen *et al.*, 2015] Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines*, pages 40–49, Beijing, China, July 2015. Association for Computational Linguistics.

[Wang, 2013] Tao Wang. Time-dependent hierarchical dirichlet model for timeline generation. *arXiv preprint arXiv:1312.2244*, 2013.

[Yan *et al.*, 2011] Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 433–443, 2011.

[Yu *et al.*, 2015] Shengkang Yu, Xi Li, Xueyi Zhao, Zhongfei Zhang, and Fei Wu. Tracking news article evolution by dense subgraph learning. *Neurocomputing*, 168:1076 – 1084, 2015.

[Yuan *et al.*, 2015] Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1351–1361, 2015.

[Zhang *et al.*, 2010] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1088. ACM, 2010.

[Zhou *et al.*, 2015] Deyu Zhou, Haiyang Xu, and Yulan He. An unsupervised bayesian modelling approach to storyline detection from news articles. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1943–1948, 2015.