An Unsupervised Bayesian Modelling Approach to Storyline Detection from News Articles

Deyu Zhou[†‡] **Haiyang Xu**[†] **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@seu.edu.cn, h.xu@seu.edu.cn, y.he@cantab.net

Abstract

Storyline detection from news articles aims at summarizing events described under a certain news topic and revealing how those events evolve over time. It is a difficult task because it requires first the detection of events from news articles published in different time periods and then the construction of storylines by linking events into coherent news stories. Moreover, each storyline has different hierarchical structures which are dependent across epochs. Existing approaches often ignore the dependency of hierarchical structures in storyline generation. In this paper, we propose an unsupervised Bayesian model, called dynamic storyline detection model, to extract structured representations and evolution patterns of storylines. The proposed model is evaluated on a large scale news corpus. Experimental results show that our proposed model outperforms several baseline approaches.

1 Introduction

The rapid development of online news media sites is accompanied by the generation of tremendous news reports. Facing such massive amount of news articles, it is crucial to develop an automated tool which can provide a temporal summary of events and their evolutions related to a topic from news reports. Therefore, storyline detection, aiming at summarising the development of certain related events, has been studied in order to help readers quickly understand the major events reported in news articles. It has attracted great attention recently. Kawamae (2011) proposed a trend analysis model which used the difference between temporal words and other words in each document to detect topic evolution over time. Ahmed et al. (2011) proposed a unified framework to group temporally and topically related news articles into same storylines in order to reveal the temporal evolution of events. Tang and Yang (2012) developed a topic-user-trend model, which incorporates user interests into the generative process of web contents. Radinsky and Horvitz (2013) built storylines based on text clustering and entity entropy to predict future events. Huang and Huang (2013) developed a mixture-event-aspect model to model sub-events into local and global aspects and utilize an optimization method to generate storylines. Wang et al. (2013) proposed an evolutionary multi-branch tree clustering method for streaming text data in which the tree construction is casted as an online posterior estimation problem by considering both the current tree and the previous tree simultaneously.

With the fast development of social media platforms, newsworthy events are widely scattered not only on traditional news media but also on social media (Zhou et al., 2015). For example, Twitter, one of the most widely adopted social media platforms, appears to cover nearly all newswire events (Petrovic et al., 2013). Therefore, approaches have also been proposed for storyline summarization on social media. Given a user input query of an ongoing event, Lin et al. (2012) extracted the storyline of an event by first obtaining relevant tweets and then generating storylines via graph optimization. In (Li and Li, 2013), an evolutionary hierarchical Dirichlet process was proposed to capture the topic evolution pattern in storyline summarization.

However, most of the aforementioned approaches do not represent events in the form of structured representation. More importantly, they ignore the dependency of the hierarchical structures of events at different epochs in a storyline. In this paper, we propose a dynamic storyline detection model to overcome the above limitations.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1943–1948,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

We assume that each document could belong to one storyline s, which is modelled as a joint distribution over some named entities e and a set of topics z. Furthermore, to link events at different epochs and detect different types of storylines, the weighted sum of storyline distribution of previous epochs is employed as the prior of the current storyline distribution. The proposed model is evaluated on a large scale news corpus. Experimental results show that our proposed model outperforms several baseline approaches.

2 Methodology

To model the generation of a storyline in consecutive time periods for a stream of documents, we propose an unsupervised latent variable model, called dynamic storyline detection model (DSDM), The graphical model of DSDM is shown in Figure 1.

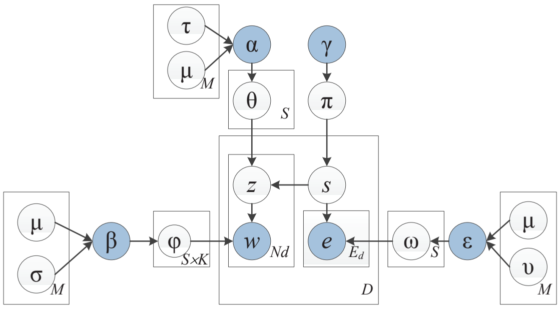


Figure 1: The Dynamic Storyline Detection model.

In this model, we assume that the storyline-topic-word, storyline-topic and storyline-entity probabilities at time t are dependent on the previous storyline-topic-word, storyline-topic and storyline-entity distributions in the last M epochs. For a certain period of time, we assume that each document could belong to one storyline s, which is modelled as a joint distribution over some named entities e and a set of topics z. This assumption essentially encourages documents published around similar time that involve the same named entities and discuss similar topics to be grouped into the same storyline. As the storyline distribution is shared across documents with the same named entities and similar topics, it essentially preserves the ambiguity that for example, documents comprising the same person and location may or may not belong to the same storyline.

The generative process of DSDM is shown below:

For each time period t from 1 to T:

- Draw a distribution over storylines $\boldsymbol{\pi}_s^t \sim$ Dirichlet(γ_s^t).
- For each storyline $s \in \{1...S\}$:
  - Draw a distribution over topics $\boldsymbol{\theta}_s^t \sim$ Dirichlet(α_s^t).
  - Draw a distribution over named entities $\boldsymbol{\omega}_s^t \sim$ Dirichlet(ϵ_s^t).
  - For each topic $k \in \{1...K\}$, draw a word distribution $\boldsymbol{\varphi}_{s,k}^t \sim$ Dirichlet(β_s^t).
- For each document $d \in \{1...D\}$:
  - Choose a storyline indicator $\boldsymbol{s}_d^t \sim$ Multinomial(π_s^t).
  - For each named entity $e \in \{1...E_d\}$:
    - Choose a named entity $\boldsymbol{e} \sim$ Multinomial(ω_s^t).
  - For other word positions $n \in \{1...N_d\}$:
    - Choose a topic $\boldsymbol{z}_n \sim$ Multinomial(θ_s^t).
    - Choose a word $\boldsymbol{w}_n \sim$ Multinomial($\varphi_{s,z}^t$).

We define an evolutionary matrix of storyline indicator s and topic z, $\sigma_{s,z,m}^t$, where each column $\sigma_{s,z,m}^t$ denotes storyline-topic-word distribution of storyline indicator s and topic z at epoch m, an evolutionary topic matrix of storyline indicator s, τ_s^t, where each column $\tau_{s,m}^t$ denotes storyline-topic distribution of storyline indicator at epoch m, an evolutionary entity matrix of storyline indicator s, υ_s^t, where each column $\upsilon_{s,m}^t$ denotes storyline-entity distribution of storyline indicator s.

We attach a vector of $M+1$ weights $\mu_{s,z}^t = \{\mu_{s,z,m}^t\}_{m=0}^M (\mu_{s,z,m}^t > 0, \sum_{m=0}^M \mu_{s,z,m}^t = 1)$, with its components representing the weights that each $\sigma_{s,z,m}^t$ contributes to calculating the priors of $\varphi_{s,z}^t$. We do it similarly for θ_s^t and ω_s^t. The Dirichlet prior for the storyline-topic-word distribution, the storyline-topic distribution and the storyline-entity distribution, respectively, at epoch t are:

$$\beta_{s,z}^t = \sum_{m=0}^{M} \mu_{s,z,m}^t \times \sigma_{s,z,m}^t \quad (1)$$

$$\alpha_s^t = \sum_{m=0}^{M} \mu_{s,m}^t \times \tau_{s,m}^t \quad (2)$$

$$\epsilon_s^t = \sum_{m=0}^{M} \mu_{s,m}^t \times \upsilon_{s,m}^t \quad (3)$$

In our experiments, the weight parameters are set to be the same regardless of storylines or topics. They are only dependent on the time window using an exponential decay function, $\mu_m =$

$\exp(-0.5 \times m)$ where m stands for the mth epoch counting backwards in the past M epochs. That is, more recent documents would have a relatively stronger influence on the model parameters in the current epoch compared to earlier documents. It is also possible to estimate the weights directly from data. We leave it as our future work.

The storyline-topic-word distribution $\varphi^t_{s,z}$, the storyline-topic distribution θ^t_s and the storyline-entity distribution ω^t_s at the current epoch t are generated from the Dirichlet distribution parameterized by $\beta^t_{s,z}, \alpha^t_s, \epsilon^t_s, \varphi^t_{s,z} \sim Dir(\beta^t_{s,z}), \varphi^t_{s,k} \sim Dir(\alpha^t_s), \omega^t_s \sim Dir(\epsilon^t_s)$. With this formulation, we can ensure that the mean of the Dirichlet parameter for the current epoch becomes proportional to the weighted sum of the word, topic distribution, and entity distribution at previous epochs.

3 Inference and Parameter Estimation

We use collapsed Gibbs sampling (Griffiths and Steyvers, 2004) to infer the parameters of the model, given observed data D. Gibbs sampling is a Markov chain Monte Carlo method which allows us 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. Letting the subscript $-d$ denote the quantity that excludes counts in document d, the conditional posterior for s_d is:

$$\begin{aligned} P(s^t_d = j|\boldsymbol{s}^t_{-d}, \boldsymbol{z}, \boldsymbol{w}, \Lambda) &\propto \frac{\{N_j\}_{-d} + \gamma}{D_{-d} + S\gamma} \\ &\times \prod_{e=1}^{E} \frac{\prod_{b=1}^{n^{(d)}_{j,e}} N_{j,e} - b + \epsilon^t_{j,e}}{\prod_{b=1}^{n^{E(d)}_j} n^E_j - b + \sum_{e=1}^{E} \epsilon^t_{j,e}} \\ &\times \prod_{k=1}^{K} \frac{\prod_{b=1}^{n^{(d)}_{j,k}} n_{j,k} - b + \alpha^t_{j,k}}{\prod_{b=1}^{n^{(d)}_j} n_j - b + \sum_{k=1}^{K} \alpha^t_{j,k}} \\ &\times \prod_{v=1}^{V} \frac{\prod_{b=1}^{n^{(d)}_{j,k,v}} n_{j,k,v} - b + \beta^t_{j,k,v}}{\prod_{b=1}^{n^{(d)}_{j,k}} n_{j,k} - b + \sum_{v=1}^{V} \beta^t_{j,k,v}}, \end{aligned}$$

where N_j denotes the number of documents assigned to storyline indicator j in the whole corpus, D is the total number of documents, $n_{j,e}$ is the number of times named entity e is assigned with storyline indicator j, n^E_j denotes the total number of named entities with storyline indicator j in the document collection, $n_{j,k}$ is the number of times words with topic label k with storyline indicator j, n_j is the total number of words (excluding named entities) in the corpus with storyline indicator j, $n_{j,k,v}$ is the number of words v with storyline indicator j and topic label k in the document collection, counts with (d) notation denote the counts relating to document d only.

Letting the index $x = (d, n)$ denote nth word in document d and the subscript $-x$ denote a quantity that excludes data from the nth word position in document d. We only sample a topic z_x if the nth word is not a named entity based on the following conditional posterior:

$$\begin{aligned} &P(z^t_x = k|s_d = j, \boldsymbol{z}_{-x}, \boldsymbol{w}, \Lambda) \\ &\propto \frac{\{n^t_{j,k}\}_{-x} + \alpha^t_{j,k}}{\{n_j\}_{-x} + \sum_{k=1}^{K} \alpha^t_{j,k}} \times \frac{\{n^t_{j,k,w_n}\}_{-x} + \beta^t_{j,k,v}}{\{n^t_{j,k}\}_{-x} + \sum_{v=1}^{V} \beta^t_{j,k,v}} \end{aligned}$$

Once the latent variables s and z are known, we can easily estimate the model parameters $\pi, \Theta, \varphi, \psi, \omega$. We set the hyperparameters $\alpha = \gamma = 0.1, \beta = \epsilon = 0.01$ for the current epoch (i.e., $m = 0$), and gather statistics in the previous 7 epochs (i.e., $M = 7$) to set the Dirichlet priors for the storyline-topic-word distribution $\varphi^t_{s,z}$, the storyline-topic distribution θ^t_s and the storyline-entity distribution ω^t_s in the current epoch t, and run Gibbs sampler for 1000 iterations and stop the iteration once the log-likelihood of the training data converges under the learned model.

4 Experiments

4.1 Dataset

We crawled and parsed the GDELT Event Database[1] containing news articles published in May 2014. We manually annotated one-week data containing 101,654 documents and identified 77 storylines for evaluation. We also report the results of our model on the one-month data containing 526,587 documents. But we only report the precision and not recall of the storylines extracted since it is time consuming to identify all the true storylines in such a large dataset. 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

[1] http://data.gdeltproject.org/events/index.html

which are verbs, nouns, or adjectives in these news articles.

4.2 Baselines

We chose the following three methods as the baseline approaches.

1. K-Means + Cosine Similarity (KMCS): the method first applies K-Means to cluster news documents for each day, then link storylines detected in different days based on the cosine similarity measurement.

2. LDA + Cosine Similarity (LDCS): the method first splits news documents on a daily basis, then applies the Latent Dirichlet Allocation (LDA) model to detect the latent storylines for the documents in each day, in which each storyline is modelled as a joint distribution over named entities and words, and finally links storylines detected in different days using the cosine similarity measurement.

3. Dynamic LDA (DLDA)[2]: this is the dynamic LDA (Blei and Lafferty, 2006) where the topic-word distributions are linked across epochs based on the Markovian assumption. That is, the topic-word distribution at the current epoch is only influenced by the topic-word distribution in the previous epoch.

4.3 Evaluation Metric

To evaluate the performance of the proposed approach, we use precision, recall and F-score 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.

4.4 Experimental Results

The proposed model is compared against the baseline approaches on the annotated one-week data which consist of 77 storylines. The number of storylines, S, and the number of topics, K, are both set to 100. The number of historical epochs, M, which is taken into account for setting the Dirichlet priors for the storyline-topic-word, the storyline-topic and the storyline-entity distributions, is set to 7. The evaluation results of our proposed approach in comparison to the three baselines are presented in Table 1.

Method	Precision	Recall	F-score
KMCS	22.73	32.47	26.74
LDCS	34.29	31.17	32.66
DLDA	62.67	61.03	61.84
DSDM	**70.67**	**68.80**	**69.27**

Table 1: Performance comparison of the storyline extraction results in terms of Precision (%), Recall (%) and F-score (%).

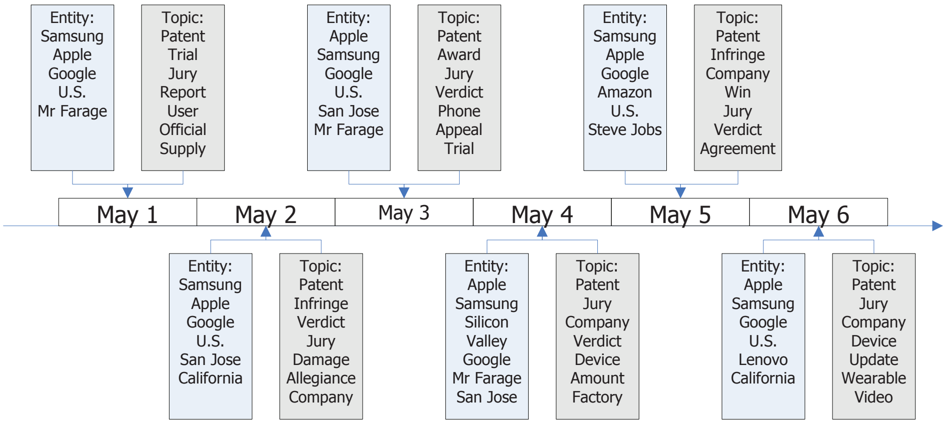
It can be observed from Table 1 that simply using K-means to cluster news articles in each day and linking similar stories across different days in hoping of identifying storylines gives the worst results. Using LDA to detect stories in each day improves the precision dramatically. The dynamic LDA model assumes topics (or stories) in the current epoch evolves from the previous epoch and further improves the storyline detection results significantly. Our proposed model aims to capture the long distance dependencies in which the statistics gathered in the past 7 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. It gives the best performance and outperforms dynamic LDA by nearly 7% in F-measure.

To study the impact of the number of topics on the performance of the proposed model, we conducted experiments on the one-month data with different number of topics varying between 100 and 200. In all these experiments, the number of storylines, S, is set to 200, based on the speculation that about 40 storylines in the annotated one-week data last for one month and about 40 new storylines occur each week. Table 2 shows the precision of the proposed method under different number of topics. It can be observed that the performance of the proposed approach is quite stable across different number of topics.

K	100	150	200
Precision	69.6%	70.2%	69.9%

Table 2: The precision of our method with various number (K) of topics.

[2] Topic number is set to 100 for both DLDA and LDCS. Cluster number is also set to 100 for KMCS.



Entity: Samsung Apple Google U.S. Mr Farage
Topic: Patent Trial Jury Report User Official Supply
Entity: Apple Samsung Google U.S. San Jose Mr Farage
Topic: Patent Award Jury Verdict Phone Appeal Trial
Entity: Samsung Apple Google Amazon U.S. Steve Jobs
Topic: Patent Infringe Company Win Jury Verdict Agreement
May 1
May 2
May 3
May 4
May 5
May 6
Entity: Samsung Apple Google U.S. San Jose California
Topic: Patent Infringe Verdict Jury Damage Allegiance Company
Entity: Apple Samsung Silicon Valley Google Mr Farage San Jose
Topic: Patent Jury Company Verdict Device Amount Factory
Entity: Apple Samsung Google U.S. Lenovo California
Topic: Patent Jury Company Device Update Wearable Video

Figure 2: Storyline about the patent infringement case between Apple and Samsung was extracted by the proposed Model.

4.5 Structured Browsing

We illustrate the evolution of storylines by using structured browsing, from which the structured information (entity, topic, keywords) about storylines and the duration of storylines can be easily observed. Figure 2 shows the storyline about "The patent infringement case between Apple and Samsung". It can be observed that in the first two days, the hierarchical structure consists of entities (Apple, Samsung) and keywords (trial, patent, infringe). The case has gained significant attention in the next three days when US jury orders Samsung to pay Apple $119.6 million. It can be observed that the stories in the next three days also consist of entities (Apple, Samsung), but with different keywords (award, patent, win). The last day's story gives an overall summary and consists of entities (Apple, Samsung) and keywords (jury, patent, company).

To further investigate the storylines detected by the proposed model, we randomly selected three detected storylines. The first one is about "the patent infringement case between Apple and Samsung". It is a short-term storyline lasting for 6 day as shown in Figure 3. The second one is about "India election", which is a long-term storyline lasting for one month. The third one is about "Pistorius shoot Steenkamp", which is an intermittent storyline, lasting for a total of 22 days but with no relevant news reports in certain days as shown in Figure 3. It can be observed that the proposed model can detect not only continuous but also intermittent storylines, which further demonstrates the advantage of the proposed model.

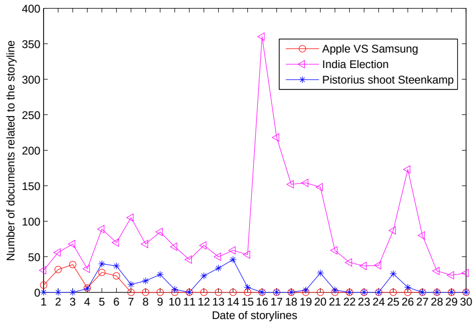


Figure 3: The number of documents on each day relating to the three storylines.

5 Conclusions and Future Work

In this paper, we have proposed an unsupervised Bayesian model to extract storylines from news corpus. Experimental results show that our proposed model is able to extract both continuous and intermittent storylines and outperforms a number of baselines. In future work, we will consider modelling background topics explicitly and investigating more principled ways in setting the weight parameters of the statistics gathered in the historical epochs. Moreover, we will also explore the impact of different scale of the dependencies from historical epochs on the distributions of the current epoch.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (61528302), State Education Ministry, the Fundamental Research Funds for the Central Universities, and the Innovate UK under the grant number 101779.

References

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

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences 101 (Suppl. 1)*, pages 5228–5235.

Lifu Huang and Lian'en Huang. 2013. 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.

Noriaki Kawamae. 2011. 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.

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

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

Saša Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton. 2013. Can twitter replace newswire for breaking news? In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*.

Kira Radinsky and Eric Horvitz. 2013. 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.

Xuning Tang and Christopher C Yang. 2012. 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.

Xiting Wang, Shixia Liu, Yangqiu Song, and Baining Guo. 2013. Mining evolutionary multi-branch trees from text streams. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 722–730. ACM.

Deyu Zhou, Liangyu Chen, and Yulan He. 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015) , Austin, Texas, USA, January 25 C 30, 2015*, pages 2468–2474.