Automatical Storyline Generation with Help from Twitter

Ting Hua
Dept. of Computer Science
Virginia Tech
tingh88@vt.edu

Xuchao Zhang
Dept. of Computer Science
Virginia Tech
xuczhang@vt.edu

Wei Wang
Dept. of Computer Science
Virginia Tech
wwang15@vt.edu

Chang-Tien Lu
Dept. of Computer Science
Virginia Tech
ctlu@vt.edu

Naren Ramakrishnan
Dept. of Computer Science
Virginia Tech
naren@cs.vt.edu

ABSTRACT

Storyline detection aims to connect seemly irrelevant single documents into meaningful chains, which provides opportunities for understanding how events evolve over time and what triggers such evolutions. Most previous work generated the storylines through unsupervised methods that can hardly reveal underlying factors driving the evolution process. This paper introduces a Bayesian model to generate storylines from massive documents and infer the corresponding hidden relations and topics. In addition, our model is the first attempt that utilizes Twitter data as human input to "supervise" the generation of storylines. Through extensive experiments, we demonstrate our proposed model can achieve significant improvement over baseline methods and can be used to discover interesting patterns for real world cases.

1. INTRODUCTION

Many philosophers such as Nietzsche believe that: nothing exists in isolation – all things are interrelated and interdependent. In the era of information explosion, although searching engines such as Google can help users reach the information of a specific event, there is still a lack of technics that can help ordinary users identify the underlying relationships between "isolated" incidents. Storyline generation is one of such technologies that give people useful insights toward better understanding of the world.

Organizing massive documents into the form of storylines can provide users with structured summaries for given subjects, showing the evolution process of relevant events. Unfortunately, detecting storylines is never an easy task. Some researchers have tried generating storylines from unorganized documents, but most of these studies were based on unsupervised clustering techniques [14, 18]. These methods can separate irrelevant storylines easily (e.g., "Sports" and "Earthquake"), however, they perform poorly in distinguishing storylines with overlapped events. As shown in Figure 1, an "Earthquake" storyline may share many common factors with a "Terrorism" storyline. For instance, both of the two storylines may involve aspects such as how many people died or injured (casualties) and how to save more lives (rescue). Previous approaches often fail to differentiate these overlapped storylines since they merely connect documents based on similarity metrics, without capturing knowledge of the hidden events and topics within the storylines.

Bayesian models such as LDA [2] are proved to be effective in learning hidden factors. Compared to clustering based approaches, few studies have been conducted in storyline generation with Bayesian networks. And the existing work often ignore the structure of storylines [4], or fail to model the hidden relations properly [20]. In this paper, we propose a hierarchical Bayesian model for Automatical Storyline Generation (ASG). As shown in Figure 1, ASG is the first storyline model with the three-level structure: storylines are root nodes, event types lie in the second-level, and the finest granularity is topic. In ASG model, different storylines can share common event types, and events can be viewed as various combinations of topics. For instance, both the storylines shown in Figure 1 include event types "Rescue" and "Casualties", and both these event types incorporate words from Topic 2 and Topic 3. ASG model also captures the relationships across the layers through two specially designed matrices. This is the first scheme by now that can quantitatively measure the hidden relations between storyline and its hidden factors.

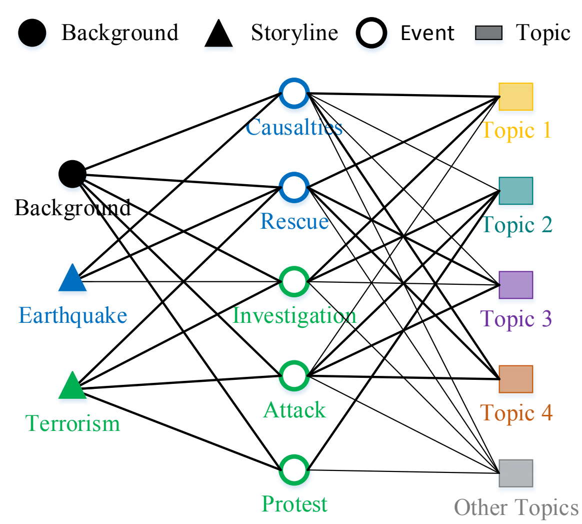

Figure 1: An example of the storyline-event-topic hierarchical structure of ASG.

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

DOI: http://dx.doi.org/10.1145/2983323.2983698

To further improve performance, ASG model uses the Twitter hashtags created by users as labels to "supervise" story generation in long news reports. Nowadays, "sharing to Twitter/Facebook" options are embedded into each news article posted on the web sites of major news media, such as CNN and BBC. When Twitter users share these articles from the original web site or retweet related tweets from their friends, they create special terms that start with #, the so-called "hashtag", to denote the topics/trends of their posts. Although Twitter data is so noisy that most existing storyline generation tools are unable to cope with it adequately [10, 15], these user self-created hashtags effectively provide human annotations for long articles. The major contributions of this paper are summarized as follows:

- **A novel Bayesian model is proposed to capture the features of real world events.** ASG model represents storyline as a three-layer structure, and provides solutions to measure hidden relations among storylines, events, and topics.
- **Human input is incorporated into the storyline generation process.** The rich up-to-date Twitter data provide the "cheapest" human made labels (hashtags), since they are publicly accessible. And ASG easily improves its performance by using these user-created Twitter hashtags to filter redundant event types.
- **An efficient Gibbs sampling inference is provided for the proposed ASG model.** Gibbs sampling was chosen for the inference and parameter estimation of ASG model for its high accuracy in estimations for LDA-like graphical model.
- **The effectiveness of the proposed ASG model is demonstrated through the comparison with existing state-of-the-art algorithms.** ASG model was tested on large datasets associated with real world events. With extensively quantitative and qualitative results, ASG model showed significant improvements over baseline methods.

2. RELATED WORK

To the best of our knowledge, this is the first attempt to generate storylines for long articles utilizing knowledge from social media, but there are several lines of related research such as topic tracking, news & Twitter modeling, and storyline discovery.

Topic tracking methods aim to identify hidden topics and track topical changes across time. Most of the earlier work in this area, for example DTM [1], estimated current topic distribution through parameters learned from the previous epoch. In addition to methods based on Markov assumptions, there has been some work modeling the evolution of topics using time stamps generated from continuous distribution [17]. TAM model [6] is a hybrid of these two approaches, which captures changes via a property dubbed a "trend class"–a latent variable with distributions over topic, words, and time. However, the granularity of "epoch" or "trend" in topic tracking approaches is inherently too fine to be suitable for the storyline discovery task.

Twitter is a newly emerging platform for news spreading [7], which covers almost all domains of newswire events [11]. Approaches that combine news and social media data in one joint model have been proposed to improve Twitter topic modeling performance, by "transferring" knowledge learned from long articles such as those in Wikipedia, blogs, news reports to short tweets [3, 5]. It is generally agreed that Twitter data is inherently noisy [19]. Therefore, few previous studies have sought to reversely use knowledge provided by social media users to label or organize long articles. Unlike these methods, our proposed ASG model only used "hashtags" from tweets, and omitted the rest of noisy content.

Storyline discovery is the closest research branch to our work. Shahaf et al. [14] proposed a metro-map format story generation framework, which first detected community clusters in each time window, and then grouped these communities into the stories. Yan et al. tracked the evolution trajectory along the timeline by emphasizing relevance, coverage, coherence and diversity of themes [18]. Mei et al. [12] proposed a HMM style probabilistic method to discover and summarize the evolutionary patterns of themes in text streams. Lappas et al. [8] designed a term burstness model to discover the temporal trend of terms in news article streams. Taking user queries as input, [10] first extracted relevant tweets and then generated storylines through graph optimization. Lin et al. [9] built a HDP (Hierarchical Dirichlet Process) model for each time epoch and then selected sentences for the storyline by considering multiple aspects such as topic relevance and coherence. Huang et al. identified local/global aspects of documents and organized these components into a storyline via optimization [4], while Zhou et al. modeled storylines as distributions over topics and named entities [20]. None of the above work jointly considered social media and news data, and all of them failed to provide a complete storyline-event-topic structure such as the one proposed in this paper.

3. MODEL

The graphical model and generative process of ASG are shown in Figure 2 and Algorithm 1 respectively. Each document d_m is a news article embed in tweet URL, associated with tweet hashtags $\mathbf{\Lambda_m}$. Storyline s is a multinational variable indicating which storyline document d_m belonging to, generated from multinomial distribution π_s. Each storyline has one multinomial distribution ψ_s over events. Variable e denotes a document's event label drawn from E-dimensional distribution ψ_s. Each event e has a multinomial distribution ϕ_e over K topics. As we will discuss in detail later, matrix ψ_s and matrix ϕ_e can reflect the relations among storyline s, event e, and topic z.

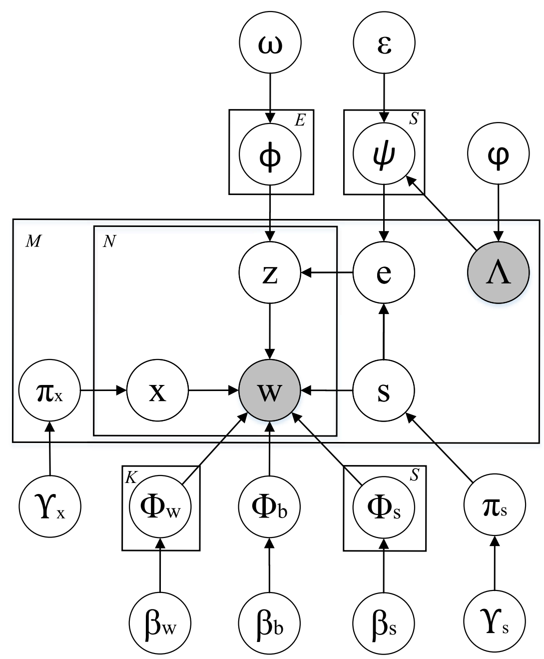

Figure 2: Graphical model for ASG.

Each word w in document d_m is associated with two labels: switching variable x and topic indicator z. $x = 0$ means the word w is generated from background distribution Φ_b, $x = 1$ shows the word is generated from storyline distribution Φ_s, and $x = 2$ denotes the word is generated via topic distribution Φ_w. Also, in the case of x taking value 2, topic z is sampled from K-dimension multinomial distribution ϕ_e. Under this strategy, ASG model can explain words in three different ways, from topics, from storylines, and from a background word distribution. The whole dataset has only one background word distribution Φ_b, while there are S different storyline-words distributions Φ_s and K different Φ_w topic distributions. This matches the intuition that a document is a mixture of background words (e.g., stop words), storyline words, and a set of aspect topics (e.g., locations).

Algorithm 1 Generation Process of ASG model

```
1:  Draw π_s ~ Dir(γ_s)
2:  Draw Φ_b ~ Dir(β_b)
3:  for each storyline s = 1, 2, ..., S do
4:      Draw Φ_s^(s) ~ Dir(β_s)
5:  for each event e = 1, 2, ..., E do
6:      Draw φ^(e) ~ Dir(ω)
7:  for each topic z = 1, 2, ..., K do
8:      Draw Φ_w^(z) ~ Dir(β_w)
9:  for each document m = 1, 2, ..., M do
10:     for each event e = 1, 2, ..., E do
11:         Draw Λ_m^(e) ~ Bernoulli(·|φ_e)
12:     Draw s ~ Multi(π_s)
13:     Generate ε_m = diag(Λ_m) × ε
14:     Draw ψ_s^(s) ~ Dir(ε_m)
15:     Draw e ~ Multi(ψ_s^(s))
16:     Draw π_x ~ Dir(γ_x)
17:     for each word w in document m do
18:         Draw x ~ Multi(π_x)
19:         Draw z ~ Multi(φ^(e))
20:         if x=0 then Draw w ~ Φ_b
21:         if x=1 then Draw w ~ Φ_s^(s)
22:         if x=2 then Draw w ~ Φ_w^(z)
```

Twitter users often create hashtags to emphasize the key points of their posts. For example, a tweet about president election may contain hashtags such as #Hillary or #Trump. As a kind of human made labels, hashtags can be used to simplify search, indexing, and topic discovery [13]. To enable modeling of context associated with hashtags, we restrict storyline-event distribution ψ_s to be filtered by its hashtags Λ_m. Namely, the choices of events for each document are restricted to its set of hashtags. Towards this goal, we set E to be the number of hashtags contained in the whole dataset. The dependency of ψ_s on both ε and Λ is shown as directed edges connecting them in Figure 2. Each element $\Lambda_m^{(e)}$ of hashtag labels Λ_m is generated through Bernoulli distribution with prior probability φ_e. As shown in Line 13 and Line 14 of Algorithm 1, storyline-event vector ψ_s is then drawn from a Dirichlet distribution with parameters $diag(\Lambda_m) \times \varepsilon$. Suppose there are 5 important hashtags in the datasets, then the value of E is 5 and Λ_m is a 5-dimension vector, within which each element is either 0 or 1. For a document associated with 2 hashtags, such as the president election tweet mentioned above, ψ_s is sampled with prior $\varepsilon_m = diag(\Lambda_m) \times \varepsilon = (0, 0, \varepsilon_3, \varepsilon_4, 0)^T$.

4. MODEL INFERENCE AND LEARNING

In this section, we first describe the inference process of collapsed Gibbs sampler for ASG mdoel, and then discuss the training and testing operations for the proposed model.

4.1 Model Inference

The prior parameter φ is d-separated from the rest of ASG model, when the hashtag labels Λ_m of document m are observed. The key to the inference is to estimate posterior distributions of hidden variables: (1) multinomial switch variable $\mathbf{x}_{mn}$ for word w_{mn}; (2) topic assignment variable $\mathbf{z}_{mn}$ for word w_{mn} when the corresponding switch variable $\mathbf{x}_{mn}$ equals to 2; (3) event assignment variable $\mathbf{e}_m$ for document m; (4) storyline assignment variable $\mathbf{s}_m$ for document m.

Gibbs sampling is chosen for the inference. First, the posterior of $\mathbf{s}_m$ is calculated through Equation (1). Due to space limitations, we only present the result here.

$$p(\mathbf{s}_m|\mathbf{w}, \mathbf{z}, \mathbf{s}_{\neg m}, \mathbf{e}, \mathbf{x}) = \prod_{v=1}^{V} \left(\frac{n_s^v + \beta_s}{\sum_{v=1}^{V} (n_s^v + \beta_s)} \right)^{n_{ms}^v} \cdot \frac{n_{se}^e + \omega}{\sum_{i=1}^{E} (n_{se}^i + \omega)} \cdot \frac{n_{ss}^i + \gamma_s}{\sum_{i=1}^{S} (n_{ss}^i + \gamma_s)} \tag{1}$$

In the above equation, V is the size of the vocabulary, E is the number of events, and S is the number of storylines. n_s^v is the number of term v choosing storyline s (when its $x = 1$), n_{ms}^v is the number of term v choosing storyline s in the scope of document m. n_{se}^i is number of document choosing event e within storyline s, and n_{ss}^i is the number of documents choosing storyline i.

The inference of $\mathbf{e}_m$ is slightly different from that of $\mathbf{s}_m$. Each storyline s has a distribution ψ_s over events. Given storyline s, document chooses corresponding event e from ψ_s:

$$p(\mathbf{e}_m|\mathbf{w}, \mathbf{z}, \mathbf{s}, \mathbf{e}_{\neg m}, \mathbf{x}) = \prod_{z=1}^{K} \left(\frac{n_e^z + \omega}{\sum_{z=1}^{K} (n_e^z + \omega)} \right)^{n_{em}^z} \frac{n_{se}^e + \varepsilon}{\sum_{i=1}^{E} (n_{se}^i + \varepsilon)}, \tag{2}$$

where K is the number of topics, n_e^z is the counts of words choosing topic z under event e, and n_{em}^z is the number of words in document m choosing topic z.

In word-level, word w_{mn} first decides its value x: (1) when $x_{mn} = 0$, word w_{mn} is sampled from background words distribution Φ_b; (2) when $x_{mn} = 1$, word w_{mn} is chosen from storyline words distribution Φ_s, where s is the choice of document m; (3) when $x_{mn} = 2$, word w_{mn} is drawn from topic distribution $\Phi_w^{(z)}$, where z is chosen beforehand by the word:

$$p(\mathbf{x}|\mathbf{w}, \mathbf{z}, \mathbf{s}, \mathbf{e}, \mathbf{x}_{\neg \mathbf{i}}) = \begin{cases} \frac{n_b^v + \beta_b}{\sum_{v=1}^{V} (n_b^v + \beta_b)} \frac{n_{xm}^0 + \gamma_x}{\sum_{i=0}^{2} (n_{xm}^i + \gamma_x)}, x = 0 \\ \frac{n_s^v + \beta_s}{\sum_{v=1}^{V} (n_s^v + \beta_s)} \frac{n_{xm}^1 + \gamma_x}{\sum_{i=0}^{2} (n_{xm}^i + \gamma_x)}, x = 1 \\ \frac{n_z^v + \beta_w}{\sum_{v=1}^{V} (n_z^v + \beta_w)} \frac{n_{xm}^2 + \gamma_x}{\sum_{i=0}^{2} (n_{xm}^i + \gamma_x)}, x = 2. \end{cases} \tag{3}$$

When $x_{mn} = 2$, topic assignment z_{mn} needs to be decided first. Similar to storyline-event relationship, each event has a distribution over topics. Given event e, topics are chosen from multinomial distribution ϕ_e:

$$p(\mathbf{z}_i|\mathbf{w}, \mathbf{z}_{\neg i}, \mathbf{s}, \mathbf{e}, \mathbf{x}) = \frac{n_k^v + \beta_w}{\sum_{v=1}^{V}(n_k^v + \beta_w)} \frac{n_e^k + \omega}{\sum_{z=1}^{K}(n_e^z + \omega)}, \quad (4)$$

where n_z^v is the number of term v choosing topic z in the scope of the whole corpus.

4.2 Learning Operations

ASG can be treated as a semi-supervised model because the hidden variables are learned under the supervision of human made hashtags. In the training process, ASG model is fed with pre-given storyline labels. That is, the storyline labels of documents can be seen by the model, while event labels, topic assignments, and switch variables are inferred to maximize the likelihood of observed words and storyline labels. With trained model $\mathcal{M}$, ASG model can be used to estimate the posterior distributions of switch variables $\tilde{\mathbf{x}}$, topics $\tilde{\mathbf{z}}$, events $\tilde{\mathbf{e}}$, and storyline labels $\tilde{\mathbf{s}}$ for new coming documents, without any pre-given labels. In order to achieve this goal, we follow the approach introduced in [16] to run the inference process on the new documents exclusively. Inference for this testing process corresponds to Equation (1)$\sim$(4) with the difference that: current Gibbs sampler is run with estimated parameters Φ_b, Φ_s, Φ_w, ϕ, ψ, and fixed hyperparameters.

Taking the inference of switch variable x for example. In the initial stage, the algorithm randomly assigns switch variables to words. Then a number of Gibbs sampling updates are conduct to estimate the posterior. Similar to switch variable $\tilde{x}$, the estimations of storyline label $\tilde{s}$, event label $\tilde{e}$, and topic assignment $\tilde{z}$ can be calculated according to Equation (1), (2), and (4).

5. EXPERIMENT

In this section, we first describe our evaluation datasets, and then compare our proposed ASG model with existing state of the-art algorithms. Finally, extensive discovery results are presented by exploring the outputs of ASG model.

5.1 Datasets and Experiment Settings

To evaluate our proposed model and other storyline generation methods, we conducted our experiments on datasets containing 110, 347 tweets and 27, 308 news articles of 6 event subjects. These events were chosen due to their great social influence and high evolution complexity. Specifically, the datasets were collected via the following steps: 1) For each event, filter Twitter data through Twitter REST API using event-relevant keywords and hashtags provided by domain experts. 2) Extract URL links embedded in the tweets, download documents associated with these links. 3) Conduct stemming and lemmatization. Note that, we didn't remove stop words in the preprocessing step, since our proposed ASG model could treat these words as background words. Table 1 lists the statistical information about evaluation datasets.

We asked human annotators to create labels for these documents: 1) select storyline label from the six event subjects for each document, and 2) assign event label from 21 event types provided by domain experts for each document. One document is associated to one storyline label and one event label. We divided the whole dataset evenly into 4 parts and assigned to 4 groups of annotators. Within each group, a label will be included in the groud-truth only if it is chosen by at least 2 out of the 3 annotators.

In our evaluation, we used weak symmetric priors for all Dirichlet parameters: $\gamma_s = 0.1$, $\gamma_x = 0.3$, $\beta_s = 0.0001$, $\beta_w = \beta_b = 0.001$, $\varepsilon = \omega = 0.01$. The number of topics K is 50, the number of

Dataset	#Tweets	#News	Country
World Cup	11K	3352	Brazil
President Election	9K	2491	Colombia
Security Protests	12K	3787	Venezuela
Education March	14K	4841	Chile
Iguala Kidnap	25K	5853	Mexico
Paris Attack	37K	6984	France

Table 1: Detailed information of datasets

events is decided by the total number of hashtags contained in the datasets, and the storyline number is set to be 6. The Gibbs sampler is run for 500 iterations with the first 100 iterations as built-in period. We compared our ASG model with the following methods. 1) Random: The random method selects documents randomly for storylines and events. 2) K-means: This method identifies storylines and events by K-means clustering. The number of clusters is set to be 6 (number of ground truth storylines). 3) LSA: LSA is a method of analyzing relationships between a set of documents and terms they contain, which uses SVD to reduce the dimension of features with high similarity. In our experiment, the number of clusters is set to be 6. 4) LDA: This method applies standard LDA twice to discovery storylines and events. First, topics found in the first run are treated as storylines ($K_1 = 6$). Then, within each obtained topic, another LDA is run to distinguish the events ($K_2 = 50$). 5) ASGH: This approach is a variant of ASG that ASGH model doesn't use Twitter hashtags for generation of events. 6) ASGB: This method is another variant of ASG that ASGB model excludes the background distribution Φ_b and has a symmetric beta prior $\gamma_x = 0.5$.

5.2 Experiment Results

Table 2 reports the ACC and NMI results for seven methods. ACC denotes the performance on storyline-level, and NMI can better evaluate the results on event-level. Random selection is the worst performer that its ratio of correct guess on storylines (ACC) is around 1/6, and the probability of successful guess in event-level (NMI) is almost equal to 0. Our ASG model outperforms the baseline methods in most of the fields, except on "World Cup", where ASGB achieves the best outcome.

Two interesting observations can be made from Table 2. First, Bayesian models perform better than clustering methods. Two clustering methods K-means and LSA are much better than random selection method, however, significantly poorer than Bayesian models LDA, ASGH, ASGB, and ASG. This is because these methods are simply built on word similarities, without further knowledge on relationship between words (hidden topics). Second, "Hashtag" factor is more important than "background words" factor. Compared to ASGH model, ASGB model obtain the performance closer to ASG model, which indicates utilization of hashtags indeed improve performance significantly. As later shown in Table 3, the scheme of background distribution can remove stop words and commonly used words, which therefore benefit the overall performance of ASG model.

Table 3 lists top 15 terms of background, storyline, and topic words learned by ASG model. In general, the ration of words assigned to background distribution, storyline distributions, and topic distributions are 33%, 27%, and 38% respectively.

We discuss the three types of words as follows. 1) **Background words**. There is only one background word distribution over the whole corpus. Most of the top ranked background words are stop words such as "en"(Spanish), "the", "de"(Spanish), or common used words such as "video", "http", "com". 2) **Storyline words**. Storyline words are not used as commonly as background words,

Table 2: Performance comparison among storyline detection methods (ACC, NMI)

	World Cup		President Election		Security Protests		Education March	
	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
Random	0.14411	2.84E-05	0.153098	0.000515	0.156086	0.00158	0.17796	0.000962
K-means	0.342921	0.02507	0.302279	0.024265	0.354775	0.16668	0.323172	0.117916
LSA	0.426756	0.000448	0.40866	0.009094	0.36727	0.117569	0.378563	0.00213
LDA	0.498904	0.280061	0.457191	0.141977	0.461266	0.271075	0.482378	0.243714
ASGH	0.512481	0.319418	0.484782	0.152756	0.517365	0.313238	0.495184	0.248761
ASGB	**0.623336**	**0.399913**	0.585973	0.181412	0.597942	0.394318	0.505529	0.268152
ASG	0.598904	0.380061	**0.622446**	**0.210056**	**0.617287**	**0.415664**	**0.528883**	**0.290283**

Table 3: Example of background/storyline/topic words learned by ASG model.

Background		Storyline 1		Storyline 2		Topic 16		Topic 21	
la	0.0093	paris	0.0104	presidente	0.0058	office	0.0076	pineda	0.0030
el	0.0087	help	0.0057	estudiante	0.0055	view	0.0074	esporte	0.0026
en	0.0058	theater	0.0049	EPN	0.0054	gobierno	0.0072	abarca	0.0025
the	0.0057	http	0.0043	news	0.0053	police	0.0061	right	0.0023
video	0.0053	shoot	0.0039	mar	0.0051	time	0.0037	zuluaga	0.0023
com	0.0051	news	0.0034	mexico	0.0049	comment	0.0034	dilma	0.0022
de	0.0038	hurt	0.0032	home	0.0047	city	0.0032	trial	0.0019
www	0.0038	hall	0.0030	iguala	0.0041	voice	0.0029	hecha	0.0018
marzo	0.0037	kill	0.0029	youtube	0.0037	new	0.0027	santos	0.0017
Twitter	0.0036	world	0.0029	gobierno	0.0035	report	0.0027	vaga	0.0016
filter	0.0034	police	0.0029	toda	0.0035	tudo	0.0027	jamariro	0.0016
apply	0.0034	terrorist	0.0026	nacional	0.0035	medium	0.0026	soho	0.0014
email	0.0033	people	0.0026	poltica	0.0035	share	0.0025	bio	0.0012
http	0.0031	maduro	0.0026	kidnap	0.0030	arrest	0.0024	bundchen	0.0011
da	0.0031	gun	0.0025	secuestrado	0.0030	agosto	0.0023	lula	0.0011

but still across a broad range of documents within the storyline. Storyline 1 seems to be related to "Paris attack", since it contains word "paris" and some other event words such as "theather", "hurt", and "shoot". Similarly, Storyline 2 is about kidnap in Mexico, which is consisted of words such as "mexico", "secuestrado" (Spanish of kidnap), and "estudiante"(Spanish of student). 3) **Topic words**. Compared to storyline words and background words, topic words are narrowed to more specifical context. For example, Topic 16 is a set of words describing government, with words such as "office", "police", and "gobierno" (Spanish of government), while Topic 21 is about people names, which includes words such as "dilma" (Brazilian president), "santos" (Colombian president), and "zuluaga" (a Colombian economist).

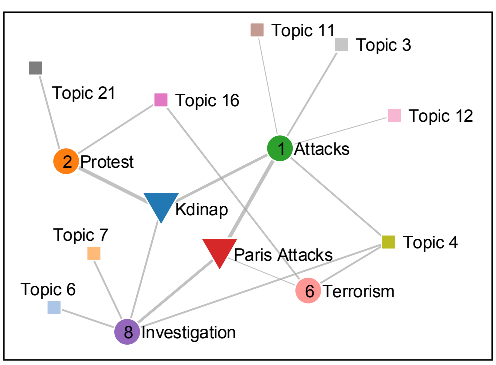

Figure 3: Relations among storyline, event type, and topics. The triangles are symbols for storylines, the circles denote event type, and the squares indicate topics.

One of the major contributions of ASG model is that: ASG is the first storyline generation model that can quantitatively measure the relations among storylines, events, and topics. In practice, these knowledge are learned from storyline-event matrix ψ and event-topic matrix ϕ. Figure 3 illustrates the usage of these matrices. Corresponding to Figure 1 in the introduction section, here triangles in the central parts denote storylines, the circles around the triangles represent event types, and the outermost squares stand for topics. The thickness of the edge between a triangle (storyline s) and a circle (event e) is proportional to the corresponding value of ψ_{se}, and similarly, the thickness of the edge between a circle (event e) and a square (topic z) is corresponding to the value of ϕ_{ez}. By correlating Table 3 with Figure 3, some interesting patterns can be obtained. 1) **Mapping ground-truth labels**. By referring to the storyline words distribution in Table 3, we mapped the storylines to real cases: the blue triangle should be the "Kidnap" storyline, with storyline words such as "mexico" and "students"; the red triangle is the "Paris attack", because words such as "terrorist" and "Paris" are top ranked in the storyline-words list. Also, as mentioned above in Table 3, topic 16 is related to "government", and topic 21 is related to "people names". 2) **Relations between topics and event types**. By referring to the values of matrix ϕ_{ez}, the result event types can be inferred through their connected topics. For example, in Figure 3, event 2 is connected to topic 16 (government) and topic 21 (people names). Among the 21 event types, the combination of the two topics is closest to event type "investigation". Similarly, event 1 is recognized as "attack", event 6 means "protest", and event 8 denotes "terrorism". 3) **Relations between storylines and event types**. The impact of event type factors to storylines can be analyzed through values of matrix ψ_{se}. As can be seen from Figure 3, both the storyline "Paris attack" and "Kidnap" are connected to event "investigation" and "attack", the difference is that: "Paris attack" storyline has a stronger connection on event "attack" over event "investigation", while storyline "Kidnap" has balanced weights on the two events. Besides the common shared events, "Paris attack" owns a private event type "terrorism", and storyline "Kidnap" has one exclusive event type "protest".

These above mentioned observations match the real world truth well, and therefore directly implies ASG model is a useful tool to identify and interpret the hidden factors that driving the evolution of events.

6. CONCLUSION

In this paper, we proposed a hierarchical Bayesian model named ASG to detect storylines. ASG is the only model with three-layer structure for the task of storyline generation. It ignores the noisy context of tweets and utilizes their promising labels made by Twitter users (hashtags) to improve the performance. Besides, through special designed data structures, ASG is capable to measure the hidden relations among storylines, events, and topics. We present the results of applying ASG model to real-word events and show its effectiveness over non-trivial baselines. Based on the outputs of ASG model, further analysis can be made to understand the underlying factors inside the documents, which can lead to a broad spectrum of future research.

7. ACKNOWLEDGEMENT

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract D12PC00337. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US government.

8. REFERENCES

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

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. In *The Journal of Machine Learning Research*, volume 3, pages 993–1022. MIT Press, 2003.

[3] Ting Hua, Ning Yue, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Topical analysis of interactions between news and social media. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2964–2971, 2016.

[4] Lifu Huang and Lian'en Huang. Optimized event storyline generation based on mixture-event-aspect model. In *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing*, pages 726–735, 2013.

[5] Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 775–784. ACM, 2011.

[6] Noriaki Kawamae. Trend analysis model: trend consists of temporal words, topics, and timestamps. In *Proceedings of the 4th ACM international conference on Web search and data mining*, pages 317–326. ACM, 2011.

[7] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[8] Theodoros Lappas, Benjamin Arai, Manolis Platakis, Dimitrios Kotsakos, and Dimitrios Gunopulos. On burstiness-aware search for document sequences. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 477–486. ACM, 2009.

[9] 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, 2013.

[10] 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.

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

[12] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 198–207. ACM, 2005.

[13] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 130–137, 2010.

[14] Dafna Shahaf, Jaewon Yang, Caroline Suen, Jeff Jacobs, Heidi Wang, and Jure Leskovec. Information cartography: creating zoomable, large-scale maps of information. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105. ACM, 2013.

[15] Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 533–542. ACM, 2013.

[16] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM, 2004.

[17] 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*, pages 424–433. ACM, 2006.

[18] Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 745–754. ACM, 2011.

[19] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916. ACM, 2014.

[20] Deyu Zhou, Haiyang Xu, and Yulan He. An unsupervised bayesian modelling approach to storyline detection from news articles. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, pages 1943–1948, 2015.