Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph

Weixin Li, Jungseock Joo, Hang Qi, and Song-Chun Zhu

Abstract—This article presents a novel method for automatically detecting and tracking news topics from multimodal TV news data. We propose a Multimodal Topic And-Or Graph (MT-AOG) to jointly represent textual and visual elements of news stories and their latent topic structures. An MT-AOG leverages a context sensitive grammar that can describe the hierarchical composition of news topics by semantic elements about people involved, related places and what happened, and model contextual relationships between elements in the hierarchy. We detect news topics through a cluster sampling process which groups stories about closely related events together. Swendsen-Wang Cuts (SWC), an effective cluster sampling algorithm, is adopted for traversing the solution space and obtaining optimal clustering solutions by maximizing a Bayesian posterior probability. The detected topics are then continuously tracked and updated with incoming news streams. We generate topic trajectories to show how topics emerge, evolve and disappear over time. The experimental results show that our method can explicitly describe the textual and visual data in news videos and produce meaningful topic trajectories. Our method also outperforms previous methods for the task of document clustering on Reuters-21578 dataset and our novel dataset, UCLA Broadcast News Dataset.

Index Terms—News topic detection and tracking, Multimodal Topic And-Or Graph, cluster sampling.

1 Introduction

1.1 Motivation and Objective

News stories provide information about real-world events and play a vital role in informing citizens, affecting public opinions and policy making. The analyses of information flow in news media, such as selection and presentation biases, agenda-setting patterns, persuasion techniques, or causal analysis are important issues in social and political science research. The primary objective of this paper is to develop an automatic *topic detection and tracking* method which can be used to analyze the real world events and their relationships.

News deals with an event and is presented in real-time as the event progresses. It updates and revises what have been reported. It also predicts the potential changes that may or may not follow in the future. Therefore, its narratives mostly focus on the temporal and causal relationships between events and how each event is dynamically transformed, based on observations made in particular points in time. Consequently, the most important thing in studying news is to understand how news stories are connected to each other over time, and this is our primal concern in this paper – to identify news stories about the same event and to monitor how they evolve.

- *W. Li and H. Qi are with the Department of Computer Science, University of California, Los Angeles (UCLA). E-mail: {lwx, hangqi}@cs.ucla.edu*
- *J. Joo is with the Department of Communication Studies, University of California, Los Angeles (UCLA). E-mail: jjoo@commstds.ucla.edu*
- *S.-C. Zhu is with the Department of Statistics and Computer Science, UCLA. E-mail: sczhu@stat.ucla.edu*

Manuscript received MM DD, YYYY; revised MM DD, YYYY.

Accordingly, we consider two related tasks in this paper: topic detection and tracking [1]. First, topic detection is aimed at clustering relevant news stories together on the fly where a topic is defined as each cluster and the corresponding multimodal model learned from it. Then we track these topics with continuously updated news data. Our objective is to generate topic trajectories to show how topics emerge, evolve, and disappear, and how their components change over time.

Our method specifically targets the domain of TV news, having two distinct properties from other types of corpora – multimodal and event-centric.

First of all, TV is a multimodal medium and TV news uses both verbal and non-verbal modalities via audio and video channels (our speech data is encoded as text via closed-captioning). Both textual and visual cues are important to understand the events described in the news. The visual dimension of mass media can be especially critical in relation to public response and engagement [2], [3]. Our model jointly captures both dimensions unlike most existing approaches in topic detection which only use text inputs.

Secondly, TV news presents stories on real-world events. For those events, the key things to understand are "who did what, when, and where." Barack Obama's winning 2008 election is a completely different event than his re-election in 2012; but they are closely related. These events dynamically introduce new people or new places involved and are eventually connected to other events. Therefore, the models to deal with TV news should be able to clearly represent the semantic structure of an event as well as its local and global changes and relations with other events.

To address these issues, we propose a novel multi-

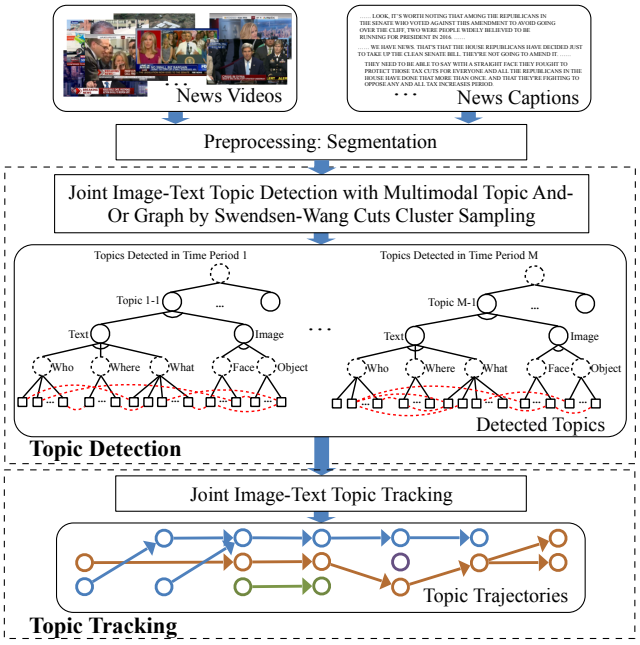


Figure 1: Overview of the proposed topic detection and tracking method. The inputs include both news videos and closed captions (texts). We detect topics through a joint image-text cluster sampling method within each time window. Then detected topics are tracked over time to form topic trajectories.

modal topic representation, i.e. Multimodal Topic And-Or Graph (MT-AOG), based on And-Or Graph (AOG), which is commonly used for various visual models [4]. The core idea of AOG is hierarchical and compositional model, which is suitable to represent the news event structures and the event relationships. To discover topics and learn the model, we also adopt a graph-partitioning based cluster sampling method, Swendsen-Wang Cuts (SWC) [5], which was originally developed for image parsing.

For evaluation, we use data from the UCLA Library Broadcast NewsScape[1], which contains a large number of broadcast news programs from the U.S. and the world since 2005. To collect the ground-truth data, we annotate a subset from the large collection. The data used in our paper will be made publicly available. Some concrete results of analysis obtained by our method including tracking the 2016 U.S. presidential election and analyzing gun shooting events can be also found at Viz2016[2].

1.2 Overview of Our Method

Fig. 1 shows an overview of our topic detection and tracking method. Both news videos and closed captions are the inputs to our method. After pre-processing steps such as story segmentation, we detect topics using a cluster sampling method, Swendsen-Wang Cuts (SWC), based on the proposed Multimodal Topic And-Or Graph (MT-AOG) which jointly models texts and images and organizes news topic components in a hierarchical structure. We further link topics detected in different time periods to generate topic trajectories which show how topics evolve over time. We describe our core representation and the main tasks in the following subsections.

1.2.1 Multimodal Topic And-Or Graph (MT-AOG)

We briefly introduce the proposed MT-AOG here. AOG has been used for modeling humans, objects and scenes in computer vision [6], [7]. MT-AOG embeds a context-sensitive grammar that jointly models hierarchical topic compositions of texts and images. There are three types of nodes in MT-AOG: AND-nodes representing compositions of sub-components (e.g. a topic is composed of the text part and the image part), OR-nodes for alternative structures (e.g. different configurations of a component in the topic structure), and TERMINAL-nodes representing the most elementary components. Fig. 2 illustrates the MT-AOG:

- *The root OR-node* in the top layer represents a number of distinct topic configurations. Each topic configuration specifies the actual contents of the topic.
- Each topic configuration is then represented by *a single topic AND-node* in the second layer. This node is composed of two parts, representing texts and images respectively.

Text Representation. The text part of each topic is represented by an *AND-node*, and its three components encode the knowledge of "who", "where" and "what." These are three key aspects in the journalism's five W's [8], [9] for describing news events and topics. More details will be provided in Section 3.2.

Image Representation. The image part of each topic is also represented by an *AND-node*. This node has two components that capture two important visual signals in news: faces and objects. Faces show the main people related to the topic, and objects include other general information about the scene and the event. More details will be shown in Section 3.3.

Joint Image-Text Representation. The relationships between image and text parts are explicitly modeled via the frequencies of pairs of an image patch and a text entity, e.g., a face and a name.

In summary, the proposed MT-AOG jointly models texts and images, and their subcomponents in a hierarchical structure. The MT-AOG model strikes a balance between the syntactic representation in natural language processing (too complex to compute) and the simplistic bag-of-words representation (too coarse). It supports the news topic detection and tracking tasks with the appropriate complexity accurately.

1.2.2 Task: Detecting and Tracking News Topics

In the massive and continuously updated news data, each news topic evolves over time. We aim to detect

1. http://newsscape.library.ucla.edu
2. http://viz2016.com

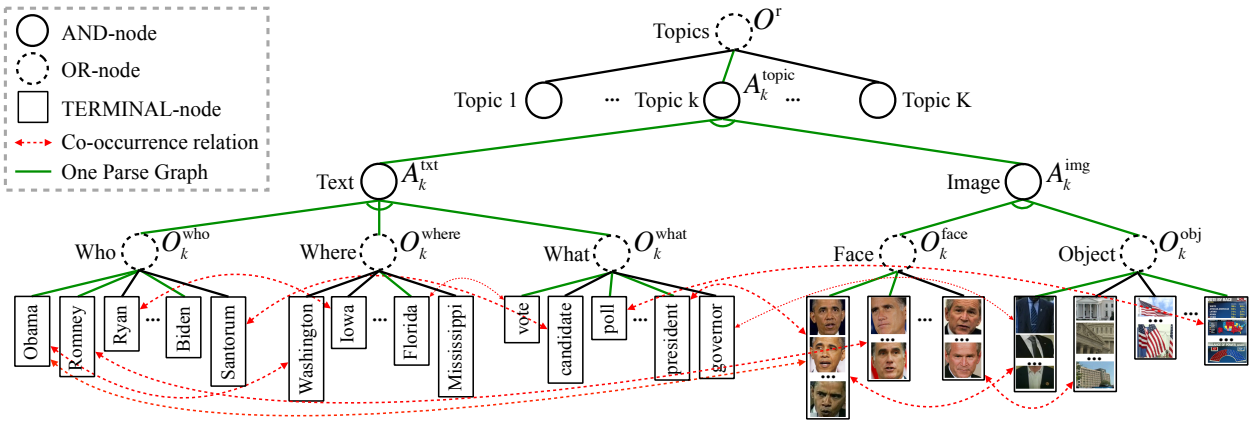


Figure 2: Illustration of our Multimodal Topic And-Or Graph. Three types of nodes are included: 1) AND-nodes representing topics' compositions, 2) OR-nodes for alternative structures, and 3) TERMINAL-nodes representing the most elementary, atomic components. The dashed red lines represent different components' co-occurring pairs. The green lines show an example of the parse graph.

topics within short time periods and further discover long-time topic trajectories. Therefore, we can show both detailed descriptions for each topic in different time periods, and how each topic develops over time. It also helps prevent the heavy computation incurred by periodically detecting topics using the entire updated news collection.

For topic detection, we group stories that elaborate the same topics. The proposed MT-AOG explicitly describes components of different topics. Thus based on the MT-AOG, we can effectively group related stories and generate meaningful topics. We solve the grouping problem using cluster sampling methods by maximizing a Bayesian posterior probability. An efficient cluster sampling algorithm introduced in image segmentation, i.e. SWC, is adopted for topic detection.

For topic tracking, we link topics detected in different time periods to generate topic trajectories. The MT-AOG model can represent topic compositions and how such information change over time. So using the MT-AOG, we can effectively track and keep updating the news states. The topics are linked by considering both topic similarities and their temporal relations.

The experimental results in Section 6 demonstrate that our method can generate meaningful topics and topic trajectories. It also achieves better performance compared to other state-of-the-art algorithms.

1.3 Summary of Contributions

This paper makes the following contributions:

- We propose a Multimodal Topic And-Or Graph that models the semantic structures of events in multimodal dimensions, which is much more suitable in the TV news domain compared to existing methods only using texts [10], [11], [12], [13], [14], [15], [16].
- We solve the topic detection problem using a cluster sampling method, Swendsen-Wang Cuts, which has better performance than commonly used greedy algorithms [11], [12], [14], [17].
- We detect and track topics simultaneously over time, generating both topic summaries in different time periods and long-time topic trajectories. The results provide useful data for further media analyses, which can hardly be fulfilled by traditional topic detection and tracking methods [1], [10].
- We propose a novel TV news dataset for joint image-text topic detection and tracking, and the ground-truth annotations for topics.

The rest of the paper is organized as follows. We first review the related literature in Section 2. Then we present the proposed topic representation in Section 3. The topic detection and tracking method are described in Section 4 and Section 5 respectively. We report our experiment results and comparisons with other state-of-the-art methods in Section 6 and conclude in Section 7.

2 Related Work

Our work is mainly related to the following four research streams: 1) topic modeling, 2) document clustering or topic detection, 3) topic tracking, and 4) news gathering and delivering systems.

1) **Topic modeling.** Probabilistic topic models [18], [19] have been widely used for detecting and analyzing latent topics, such as the latent Dirichlet allocation (LDA) model [11], [20] and its extensions [21], [22], [23]. Even though these methods are effective in general topic modeling, they typically rely on the bag-of-words (BoW) representation. The BoW representation is computationally efficient, but it ignores the semantic and compositional structures of news events. News stories are generally driven by events, so information from aspects like "who", "where" and "what" is crucial for summarizing these stories and generating meaningful news topics. Newman et al. [24] considered these aspects but included them as a whole. Li et al. [25] used

this information but assume that these components are independent. Moreover, all the aforementioned methods are unimodal methods which only use texts.

Multimodal probabilistic topic models have also been proposed in the literature [26], [27], [28]. To detect Twitter events, Cai et al. [29] proposed a Spatial-Temporal Multimodal TwitterLDA model which uses five Twitter cues including text, image, location, timestamp, and hashtag, and modeled topics as location-specific distributions. Qian et al. [30] proposed a multimodal event topic model for social event analysis. But in their model, no compositional structures are considered for the textual or visual modality. Chen et al. [31] proposed a Visual-Emotional LDA (VELDA) model which relates tweet images and texts both visually and emotionally for image retrieval. Jia et al. [32] proposed a Multimodal Document Random Field (MDRF) model for image retrieval, which is built using a Markov random field over LDA. For both VELDA and MDRF, there is only one image for one document. Our method is designed to detect and track news topics using broadcast news videos.

We pose the topic detection problem as a graph partitioning problem, and organize news stories in a graph. Some probabilistic topic models also build document networks. The Rational Topic Model (RTM) proposed by Chang et al. [33], and Semi-Supervised Relational Topic Model (ss-RTM) proposed by Niu et al. [34] are both extensions of LDA which account for links between documents when modeling topics. RTM models networks of text data, e.g. citation networks of documents. Ss-RTM is designed for recognizing images with text tags in social media. It jointly models image contents and their links (two images are linked if they share one or more common text tags). Both RTM and ssRTM use data from one modality to build links, and use data from another modality in nodes, while our method jointly models both texts and visuals in nodes and links. Our model for graph partitioning also considers the total partition number and partition size distributions (see Section 3 and 4 for more details).

2) **Document clustering or topic detection.** Clustering based methods are also widely used for the task of news topic detection. A large number of methods for topic detection in the Topic Detection and Tracking (TDT) research [1] (e.g. [10], [35]) use clustering methods for detecting news topics, where stories on the same topic are gathered. Traditional document clustering methods [36], [37] can also be used for topic detection. However, most of these methods work on unimodal data and mainly focus on the text domain.

Multimodal topic clustering methods have been proposed by taking both texts and visuals into consideration. In most of these methods, texts are represented using the BoW representation [14], [17], [38], [13]. For visual representation, some methods use color histograms of the keyframes [14]. Other methods detect the near-duplicate keyframes (NDK) first and then use them to build visual relations between news stories [17], [38]. Even though these methods can compute the visual similarities between stories, they are not capable of modeling the decomposition of visual parts in news topics. In terms of the clustering methods, [14] and [17] used co-clustering algorithm and one of its extensions with constraints added respectively. [13] groups news stories based on the linear combination of textual and visual similarities. [38] detects topics within one multimodal graph, which is obtained by merging one text graph and another visual graph constructed based on LDA and NDK respectively.

Some work also combined topic modeling and document clustering together, such as the multi-grain clustering topic model (MGCTM) proposed by Xie et al. [12]. They showed that these two tasks are closely related and can help each other as both performances are improved. This work still remains in the pure text domain and uses the BoW representation.

3) **Topic tracking.** The traditional topic tracking problem in TDT ([1], [10]) is defined as the process of finding related additional stories for some pre-learned topics. Many methods have been proposed for solving this problem such as those in [1], [39], [40]. However, deciding the topic of each incoming story based on the previous learned topics can take a long time in a large data collection.

In the probabilistic modeling community, some models incorporate time information, such as the Dynamic Topic Model (DTM) [21] which models topic evolution over time, and the temporal Dirichlet mixture model (TDPM) [41] for evolutionary clustering. In DTM, it is assumed that topics exist throughout the whole time period, which is usually not the case in the news domain. TDPM generates clusters that fit the data during each time period as much as possible while preserves the smoothness of clustering results over time. Both DTM and TDPM are unimodal.

Instead of using the previous two methods, we choose to do topic tracking by linking topics detected in different time periods. Some linking methods, such as those by Mei et al. [42] and Kim et al. [43], are closely related to our topic tracking task. However, the method in [42] is designed for news about some specific topics such as "tsunami." The similarity matrices used in [43] are based on the topics obtained by the original LDA model with BoW assumption. Moreover, both of the two methods are based on textual information only.

4) **News gathering and delivering system.** Several news gathering and delivering systems have been presented recently, such as News Rover [44], [45] and EigenNews [46], [47]. News Rover relies on external sources (e.g. Google News, which presumably uses user-click data, etc.) to get corresponding topics for TV news stories. TV news stories and collected topics are linked using the combination of NDK based visual similarity and BoW based textual similarity. EigenNews focuses on individual stories without the notion of topic. It

discovers links among news stories and online articles by matching keyframes based on local visual features or matching texts based on BoW histograms and named entities. Different from these two systems, we learn topics solely from TV news data. Another difference of our method is that we use a joint probabilistic model of images and texts, and perform learning and inference on this unified representation.

Besides the previous four research streams, our work is also related to event coreference resolution [48], [49]. Zhang et al. [49] proposed to detect coreferential news event pairs by incorporating textual and visual similarities. However, coreferential events are defined to be the specific event occurrence mentioned in different sentences/documents with exactly the same characteristics (location, time, involved people, etc.), so event coreference resolution is not designed to deal with event evolutions, which is the goal of this paper.

Since we use entities in the topic representation (i.e. "who", and "where"), our work is also related to another problem in the literature: Knowledge Base Population (KBP), which is the task of discovering facts about entities to augment a knowledge base [50], [51]. Different from the KBP problem, our work focuses on detecting and tracking topics using entities as features of news stories.

In this paper, we collect a new dataset named UCLA Broadcast News Dataset since there is a lack of publicly available multimedia dataset for news topic detection and tracking. Even though some multimedia news datasets have been used in previous work, such as the Topic Detection and Tracking (TDT) datasets [1], and the TRECVID corpus [52], they are not publicly available, and some of them do not have ground-truth annotations. News video datasets for other tasks have also been presented in the literature, e.g. the REPERE corpus for multimodal person recognition [53] and Stanford I2V dataset for image-to-video visual search [54].

3 Topic Representation

In this section, we define our Multimodal Topic And-Or Graph (MT-AOG) for topic representation.

3.1 Overall Representation

A **MT-AOG** can be defined by a three-tuple $\mathcal{G} = (V, E, \Theta)$. The node set V consists of three subsets of nodes: **AND-nodes** V_{AND}, **OR-nodes** V_{OR} and **TERMINAL-nodes** V_{T}, i.e. $V = V_{\mathrm{AND}} \cup V_{\mathrm{OR}} \cup V_{\mathrm{T}}$. E denotes the edge set in the graph. Θ represents the MT-AOG model parameters. We have $\Theta = \{K, \theta_1, ..., \theta_K\}$ where K is the total topic number, and $\theta_1, ..., \theta_K$ represent the model parameters for these K topics respectively. Fig. 2 illustrates the proposed MT-AOG topic representation.

A **parse graph** pg is an instantiation of the MT-AOG by selecting children nodes at OR-nodes (according to the scoring functions defined below in this Section and Section 3.2, 3.3, and 3.4). The green lines in Fig. 2 shows one example of the parse graph.

As shown in Fig. 2, the MT-AOG has five layers. Nodes in each layer are explained as follows:

1) **Root OR-node** $O^{\mathrm{r}} \in V_{\mathrm{OR}}$ in the first layer of MT-AOG represents different topic configurations and their mutual contextual information. Each topic k ($k = 1, ..., K$) is represented by an AND-node A_k^{topic} in the second layer with the model parameter θ_k.

News stories are reports of topics, i.e. topic instances, from various TV news networks. To find the optimal pg for one news story, i.e. the optimal topic instantiation of the MT-AOG, we define a series of scoring functions at different nodes below. We denote a news story by $\mathbf{d}_i$. For a story $\mathbf{d}_i$, the scoring function at root OR-node O^{r} is defined as:

$$s^{\mathrm{root}}(\mathbf{d}_i; \Theta) = \max_{\theta_k \in \Theta} s^{\mathrm{topic}}(\mathbf{d}_i; \theta_k), \tag{1}$$

where $s^{\mathrm{topic}}(\mathbf{d}_i; \theta_k)$ is the scoring function at A_k^{topic}, which will be introduced later. In the following sections, we omit the story index, i, for simplicity.

2) **Topic AND-node** $A_k^{\mathrm{topic}} \in V_{\mathrm{AND}}$ represents one topic configuration. One topic is composed of the text part and the image part. So A_k^{topic} has two children AND-nodes, i.e. text AND-node A_k^{txt} and image AND-node A_k^{img}. Considering both text and image parts and their contextual relations, we define the scoring function at A_k^{topic} as:

$$\begin{aligned} s^{\mathrm{topic}}(\mathbf{d}; \theta_k) = {} & s^{\mathrm{txt}}(\mathbf{d}^{\mathrm{txt}}; \theta_k) + s^{\mathrm{img}}(\mathbf{d}^{\mathrm{img}}; \theta_k) \\ & + s^{\mathrm{joint}}(\mathbf{d}^{\mathrm{joint}}; \theta_k) + g(f_{A_k^{\mathrm{topic}}}), \end{aligned} \tag{2}$$

where $\mathbf{d}^{\mathrm{txt}}$, $\mathbf{d}^{\mathrm{img}}$ and $\mathbf{d}^{\mathrm{joint}}$ denote the text part, the image part and their joint information of the story $\mathbf{d}$ respectively ($\mathbf{d} = \mathbf{d}^{\mathrm{txt}} \cup \mathbf{d}^{\mathrm{img}} \cup \mathbf{d}^{\mathrm{joint}}$). The two terms $s^{\mathrm{txt}}(\mathbf{d}^{\mathrm{txt}}; \theta_k)$ and $s^{\mathrm{img}}(\mathbf{d}^{\mathrm{img}}; \theta_k)$ are scoring functions at A_k^{txt} and A_k^{img} respectively. The term $s^{\mathrm{joint}}(\mathbf{d}^{\mathrm{joint}}; \theta_k)$ describes the contextual relations between the text part and the image part. These three terms will be explained later. To take the prior of choosing A_k^{topic} at root node O^{r} into consideration, we also add function $g(f_{A_k^{\mathrm{topic}}})$ in the scoring function, where $f_{A_k^{\mathrm{topic}}} \in \theta_k$ is the branching frequency. We observed that in broadcast news, dominant topics with a large amount of coverage are only a small part of all the topics, and sizes of most topics are small. Accordingly, we assume that branching frequencies at O^{r} follow a power law distribution[3] (the verification of our observation will be shown in Section 3.5).

3.2 Text Representation

For a news story $\mathbf{d}$, its *text part* $\mathbf{d}^{\mathrm{txt}}$ contains the "who" component $\mathbf{d}^{\mathrm{who}}$, the "where" component $\mathbf{d}^{\mathrm{where}}$, and

3. In the experiments, for the function $g(\cdot)$, we use the Zipf's law probability distribution, i.e. $g(f) = \frac{f^{-\rho}}{\zeta(\rho)}$ and set the parameter ρ that describes the distribution's exponent as $\rho = 1.75$ (ζ is the Riemann Zeta function).

the "what" component $\mathbf{d}^{\text{what}}$. These three components describe the people involved, related places, and what happened respectively. We extract words for different components by performing named entity extraction using the Stanford Named Entity Recognizer [55]. Thus each of these three components can be represented by a list of words (word duplication is allowed in the list), e.g. $\mathbf{d}^{\text{who}} = (w_1, ..., w_{M^{\text{who}}})$ where M^{who} is the total word number in the "who" component in story $\mathbf{d}$. The total numbers of words in "where" and "what" components are denoted by M^{where} and M^{what} respectively.

We extract co-occurring word pairs from the three components in the text part. We consider a pair of words as one co-occurring pair if they belong to two different components, and are extracted from the same sentence. The list of co-occurring word pairs in $\mathbf{d}$ is denoted by $\mathbf{d}^{\text{pair}}_{\text{txt}} = [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{who}}, w_2 \in \mathbf{d}^{\text{where}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{who}}, w_2 \in \mathbf{d}^{\text{what}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{where}}, w_2 \in \mathbf{d}^{\text{what}}]$.

Text AND-node A^{txt}_k in the MT-AOG has three children OR-nodes, i.e. O^{who}_k, O^{where}_k, and O^{what}_k, which represent "who", "where" and "what" components in the text part of topic k respectively. To model these three components jointly, the scoring function at A^{txt}_k (i.e. $s^{\text{txt}}(\mathbf{d}^{\text{txt}}; \theta_k)$ in Eq. 2) is defined as:

$$s^{\text{txt}}(\mathbf{d}^{\text{txt}}; \theta_k) = \sum_{c \in \{\text{who,where,what}\}} s^{\text{c}}(\mathbf{d}^{c}; \theta_k) + s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{txt}}; \theta_k), \tag{3}$$

where the variable c represents the component type $c \in \{\text{who}, \text{where}, \text{what}\}$. $s^{\text{c}}(\mathbf{d}^{c}; \theta_k)$ represents the scoring function at the OR-node for one component $O^{c}_k \in \{O^{\text{who}}_k, O^{\text{where}}_k, O^{\text{what}}_k\}$. $s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{txt}}; \theta_k)$ describes the contextual relations between components in the text part:

$$s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{txt}}; \theta_k) = \sum_{c_1, c_2} \mathcal{N}(\frac{M^{c_1}}{M^{c_2}}; \mu^{c_1 c_2}_k, \sigma^{c_1 c_2}_k) + \sum_{(w_1, w_2) \in \mathbf{d}^{\text{pair}}_{\text{txt}}} \log(f^{(w_1, w_2)}_k + 1), \tag{4}$$

where we have the component types $c_1 \in \{\text{who}, \text{where}\}$, $c_2 \in \{\text{where}, \text{what}\}$, $c_1 \neq c_2$. $\frac{M^{c_1}}{M^{c_2}}$ represents the ratio of word numbers from two different components. The three ratios, namely $\frac{M^{\text{who}}}{M^{\text{where}}}$, $\frac{M^{\text{who}}}{M^{\text{what}}}$, and $\frac{M^{\text{where}}}{M^{\text{what}}}$ are assumed to follow Gaussian distributions $\mathcal{N}(\frac{M^{c_1}}{M^{c_2}}; \mu^{c_1 c_2}_k, \sigma^{c_1 c_2}_k)$. $\mu^{c_1 c_2}_k, \sigma^{c_1 c_2}_k \in \theta_k$ are parameters for corresponding Gaussian distributions. The parameter $f^{(w_1, w_2)}_k \in \theta_k$ is the frequency of co-occurring word pair (w_1, w_2) in topic k.

Three children OR-nodes of A^{txt}_k in the fourth layer, namely O^{who}_k, O^{where}_k, and O^{what}_k, describe a set of possible words for the corresponding components. A certain news story may trigger a subset of these words. The words are represented by TERMINAL-nodes in the last layer. The scoring functions at these OR-nodes are defined as:

$$s^{\text{c}}(\mathbf{d}^{c}; \theta_k) = \sum_{w \in \mathbf{d}^{c}} \log(f^{w}_k + 1) \tag{5}$$

Figure 3: A common example pair of a face and a object cluster discovered by our algorithm.

where the component type $c \in \{\text{who}, \text{where}, \text{what}\}$. The parameter $f^{w}_k \in \theta_k$ is the frequency of word w in topic k.

3.3 Image Representation

The story's *image part* $\mathbf{d}^{\text{img}}$ contains the face component $\mathbf{d}^{\text{face}}$, and the object component $\mathbf{d}^{\text{obj}}$, i.e. $\mathbf{d}^{\text{face}}, \mathbf{d}^{\text{obj}} \in \mathbf{d}^{\text{img}}$. Each entity in the face/object component corresponds to one cluster of face/object patches. To obtain the face component, we first perform face detection using the Viola-Jones face detector [56] and extract face features based on Local Binary Pattern [57] and Local Gabor Binary Pattern Histogram Sequence [58], and then use the k-means algorithm to cluster faces into groups. To get the object components, we first extract patches from images using Selective Search [59] which generates object proposals. We then extract a 4096-dimensional feature vector for each patch from the fc7 layer of AlexNet [60] trained on ImageNet data [61]. We use a pretrained model in Caffe [62]. Then we cluster these patches by k-means algorithm. Fig. 3 illustrates how one image can be parsed based on the obtained face and object clusters. Each face/object patch can be represented by its corresponding cluster membership. Then the face and object components of one story $\mathbf{d}$ can also be represented by a list of visual words, e.g. $\mathbf{d}^{\text{face}} = (w_1, ..., w_{M^{\text{face}}})$ where each word $w_j \in \mathbf{d}^{\text{face}}$ represent one face patch's cluster membership. M^{face} is the total number of face patches in $\mathbf{d}^{\text{img}}$ and the total number of object patches is denoted by M^{obj}.

We extract co-occurring word pairs from the face and object components of the image part. A pair of visual words is considered as one co-occurring pair if the two words are from the face and object components respectively, and they both appear in one short time period in the news video. We denote the list of co-occurring pairs extracted from the image part by $\mathbf{d}^{\text{pair}}_{\text{img}} = [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{face}}, w_2 \in \mathbf{d}^{\text{obj}}]$.

Image AND-node A^{img}_k in the MT-AOG represents the image part of topic k. It has two children OR-nodes, i.e. O^{face}_k and O^{obj}_k, which represent face and object components respectively. The scoring function at A^{img}_k is defined in a similar way to the one at A^{txt}_k:

$$s^{\text{img}}(\mathbf{d}^{\text{img}}; \theta_k) = \sum_{c \in \{\text{face,obj}\}} s^{\text{c}}(\mathbf{d}^{c}; \theta_k) + s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{img}}; \theta_k), \tag{6}$$

where the component type $c \in \{\text{face}, \text{obj}\}$. The term $s^{\text{c}}(\mathbf{d}^c; \theta_k)$ represents the scoring function at OR-node for one component $O_k^c \in \{O_k^{\text{face}}, O_k^{\text{obj}}\}$.

The term $s^{\text{pair}}(\mathbf{d}_{\text{img}}^{\text{pair}}; \theta_k)$ describes contextual relations between face and object components and we define it as:

$$s^{\text{pair}}(\mathbf{d}_{\text{img}}^{\text{pair}}; \theta_k) = \sum_{(w_1,w_2)\in \mathbf{d}_{\text{img}}^{\text{pair}}} \log(f_k^{(w_1,w_2)} + 1), \quad (7)$$

where $f_k^{(w_1,w_2)} \in \theta_k$ is the frequency of co-occurring visual word pair (w_1, w_2) in the topic k.

Two children OR-nodes of A_k^{img}, namely O_k^{face}, and O_k^{obj}, can describe a set of alternative visual words. These words are represented by TERMINAL-nodes in the last layer. Scoring functions at these OR-nodes, i.e. $s^{\text{c}}(\mathbf{d}^c; \theta_k), c \in \{\text{face}, \text{obj}\}$, are defined in the same way as those at O_k^{who}, O_k^{where} and O_k^{what} (Eq. 5).

3.4 Joint Image-Text Representation

To jointly model the topic text and image parts, we extract their co-occurring word pairs. Three kinds of pairs, namely the face-who, face-what, and object-what pairs, are obtained for each news story $\mathbf{d}$. The words in each co-occurring pair appear in one short time period. These image-text co-occurring word pairs are denoted by $\mathbf{d}^{\text{joint}} = [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{who}}, w_2 \in \mathbf{d}^{\text{face}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{what}}, w_2 \in \mathbf{d}^{\text{face}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{what}}, w_2 \in \mathbf{d}^{\text{obj}}]$. We use M^{txt} and M^{img} to denote the total entity numbers of the text part and the image part in $\mathbf{d}$ respectively. So we have $M^{\text{txt}} = M^{\text{who}} + M^{\text{where}} + M^{\text{what}}$, and $M^{\text{img}} = M^{\text{face}} + M^{\text{obj}}$.

The score function $s^{\text{joint}}(\mathbf{d}^{\text{joint}}; \theta_k)$ in Eq. 2 is defined as:

$$s^{\text{joint}}(\mathbf{d}^{\text{joint}}; \theta_k) = \mathcal{N}(\frac{M^{\text{txt}}}{M^{\text{img}}}; \mu_k^{\text{joint}}, \sigma_k^{\text{joint}}) + \sum_{(w_1,w_2)\in \mathbf{d}^{\text{joint}}} \log(f_k^{(w_1,w_2)} + 1). \quad (8)$$

We assume that the ratio between the total entity numbers of the text part and the image part, i.e. $\frac{M^{\text{txt}}}{M^{\text{img}}}$, follows Gaussian distribution $\mathcal{N}(\frac{M^{\text{txt}}}{M^{\text{img}}}; \mu_k^{\text{joint}}, \sigma_k^{\text{joint}})$ with the parameters $\mu_k^{\text{joint}}, \sigma_k^{\text{joint}} \in \theta_k$. The parameter $f_k^{(w_1,w_2)} \in \theta_k$ is the frequency of the word pair (w_1, w_2) in topic k.

Based on the previous scoring functions, we can select the best children nodes at OR-nodes and find the optimal parse graph pg^* for the story $\mathbf{d}$ by calculating $s^{\text{root}}(\mathbf{d}; \Theta)$.

3.5 Empirical Evaluations of Assumptions in MT-AOG

In the MT-AOG representation, we make the assumption that branching frequencies at root OR-node O^{r} follow the power law distribution. To verify our assumption, we collected a news corpus that contains 1,853 news stories during a period of seven days. Annotators were asked to group the stories according to their topics and we collected 355 topics in total after annotation.

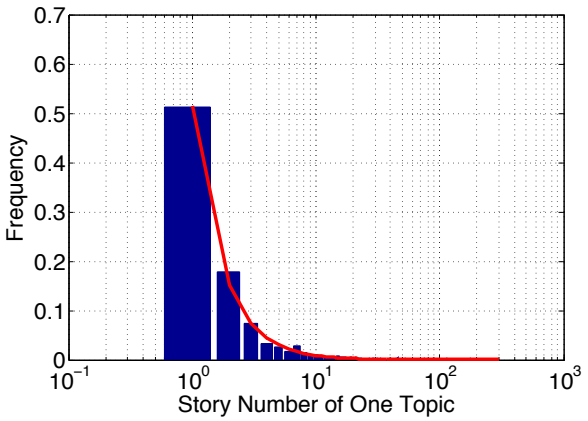


Figure 4: The histogram of the number of stories in each topic and the fitting result (the red curve).

To verify that the branching frequencies at O^{r} follow the power law distribution, using the collected corpus, we fit the empirical distribution of the story numbers in the topics to the power law distribution. The p-value (at the 5% significance level) is 0.9984. Fig. 4 shows the empirical distribution and the fitted curve (red line).

4 Topic Detection

In this section, we present our formulation of the topic detection problem, and the algorithm for optimizing a Bayesian posterior probability for the problem.

4.1 Problem Formulation

With the MT-AOG topic representation, our goal of topic detection is to cluster news stories that describe the same topics and obtain the MT-AOG model parameters Θ for the topics. We pose this clustering problem as a graph partitioning problem in which news stories, as vertices in the adjacency graph, are partitioned into coherent groups. We show one example of the adjacency graph in Fig. 5. Edges in the adjacency graph are associated with certain weights corresponding to related story similarities. Partitions can be obtained by dividing the vertices into groups with specific properties and also keeping the number of edges between separated components small. Graph partitioning can help the news topic detection since even though news stories from one topic develop over time and drift the topic, they can still be grouped together through the connections between temporally adjacent stories with less changes and more similarities.

Formally, we are given a news story corpus that contains N news stories, i.e. $D = \{\mathbf{d}_i; i = 1, \ldots, N\}$. The adjacency graph is defined as $\mathcal{G}_{\text{ADJ}} = (V_{\text{ADJ}}, E_{\text{ADJ}})$ where V_{ADJ} is a set of vertices and each vertex $v_i \in V_{\text{ADJ}}$ corresponds to one news story $\mathbf{d}_i$. E_{ADJ} is a set of edges between vertices. The clustering/partition W we are trying to find given D is defined as:

$$W = (K, \pi_K, \Theta), \quad (9)$$

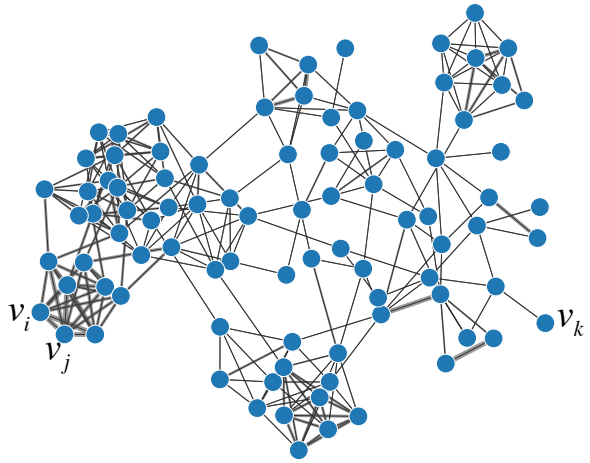


Figure 5: One adjacency graph. Each vertex in the graph corresponds to one news story. Edges are associated with weights corresponding to the story similarities (the edge thickness shows the story similarity). The vertices v_i and v_j both talk about the Oklahoma tornado topic and they are adjacent to each other in the graph. The other vertex v_k which is far away from v_i and v_j talks about the California High-Speed Rail project.

where K is determined automatically while solving the partitioning problem and π_K represents the $K-$partition of the adjacency graph. π_K is defined as:

$$\pi_K = (V_1, ..., V_K), \bigcup_{k=1}^{K} V_k = V_{\text{ADJ}}, V_k \cap V_j = \emptyset, \forall i \neq j. \tag{10}$$

This becomes an optimization problem which can be solved by maximizing a Bayesian posterior probability:

$$W^* = \arg\max_{W \in \Omega} p(W|D) = \arg\max_{W \in \Omega} p(D|W)p(W), \tag{11}$$

where Ω is the solution space. The likelihood probability $p(D|W)$ is formulated as:

$$p(D|W) = \prod_{i=1}^{N} p(\mathbf{d}_i; \Theta) \propto \exp\{\sum_{d_i \in D} s^{\text{root}}(\mathbf{d}_i; \Theta)\}. \tag{12}$$

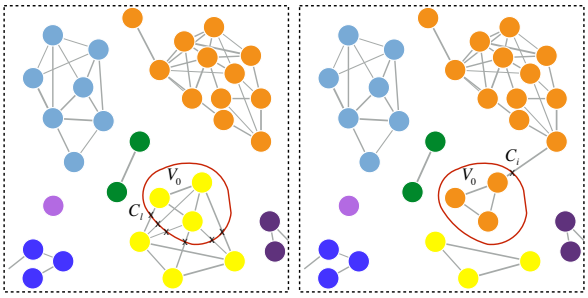
The prior probability $p(W)$ penalizes the partition number K in W and we formulate it as:

$$p(W) \propto \exp\{-\alpha N K\}. \tag{13}$$

α is a positive parameter which acts as a threshold for grouping stories into topics. This prior helps us combine close partitions to get dense results.

4.2 Inference by Swendsen-Wang Cuts

To solve the topic detection problem formulated above, we adopt a cluster sampling method Swendsen-Wang Cuts (SWC) [5]. It is a Markov Chain Monte Carlo method which samples the solution space Ω efficiently. An alternative method will be the expectation-maximization (EM) algorithm. But in [63], SWC is shown to be more effective than EM which finds only a local minimum.



(a) Before flipping. (b) After flipping.

Figure 6: SWC flips the selected component V_0. The cuts are marked with crosses.

SWC changes the labels of a group of vertices at the same time. It thus solves the coupling problem of Gibbs sampler (which flips a single vertex) by quickly jumping between local minima. SWC starts with an initial partition π, which can be the one which sets all stories to be in the same group, or can be set randomly. We denote the set of edges whose related two vertices belong to the same group under the partition π by $E(\pi)$. The optimal clustering W^* can be obtained by performing the following steps iteratively until convergence.

(1) *Determining edge status.* Each edge $e =< v_i, v_j > \in E(\pi)$ is associated with a Bernoulli random variable $u_e \in \{0, 1\}$ which indicates the edge's on/off status and a turn-on probability q_e. We define:

$$q_e = e^{-\mathcal{D}(e)/T}, \tag{14}$$

where T is the temperature used in the simulated annealing procedure and it is slowly decreased according to a cooling schedule. $\mathcal{D}(e)$ is the distance of these two vertices obtained based on the Kullback-Leibler (KL) divergence:

$$\mathcal{D}(e) = \sum_{F \in \mathcal{F}} \lambda_F \cdot \frac{KL(F(v_i)||F(v_j)) + KL(F(v_j)||F(v_i))}{2}, \tag{15}$$

where $F(\cdot)$ denotes one type of feature of the vertex and λ_F is the weight for feature F. Here we use the distributions for the five components in the text and image parts (i.e. who, where, what, face and object) to construct the feature set $\mathcal{F}$. Moreover, since KL divergence is non-symmetric, we average the KL divergence of $F(v_i)$ given $F(v_j)$ and the KL divergence of $F(v_j)$ given $F(v_i)$ to get a symmetric distance measure for vertices v_i and v_j. Based on these definitions, in this step, we set $u_e = 0$ (i.e, turn e off) with probability $1 - q_e$ for all $e \in E(\pi)$.

(2) *Computing connected components.* Once the states u_e is determined for each edge $e \in E(\pi)$, the graph G is partitioned into a set of connected components, each of which contains vertices that belong to the same group.

(3) *Selecting a component and flipping it.* Among all the connected components formed in (2), we can randomly select one component V_0 to flip. We show one example of V_0 in Fig. 6a. The target label for V_0 can be a new one

that has not been used yet or just the same as any other connected components, thus allowing reversible jumps in the solution space. The current partition number is denoted as K'. Then the number of possible new labels for the selected component is $K'+1$. Assuming that $V_0 \subseteq V_l$ in the current partition π, we denote a series of sets

$$S_1 = V_1, S_2 = V_2, ..., S_l = V_l \backslash V_0, S_{K'} = V_{K'}, S_{K'+1} = \emptyset \tag{16}$$

that V_0 can be merged with. Then V_0 can be flipped by drawing a random sample l' with probability

$$p(l'(V_0) = i | V_0, \pi) = \frac{\gamma_i p(\pi_i|D)}{\sum_{j=1}^{K'+1} \gamma_j p(\pi_j|D)}, \tag{17}$$

where π_i is the partition after assigning the label of the component V_0 to be i and keeping other components' labels the same as in π. We also have

$$\gamma_i = \prod_{e \in \mathcal{C}_i} (1 - q_e), \tag{18}$$

where $\mathcal{C}_i$ is the cuts between V_0 and S_i, i.e. $\mathcal{C}_i = \mathcal{C}(V_0, S_i) = \{< s, t >: s \in V_0, t \in S_i\}$. Two examples of the cuts are shown in Fig. 6, which are marked by the crosses. Theorem 3 in [5] proved that the acceptance rate will be 1 by choosing the new label of V_0 by Eq. 17.

Another thing to be noted is that when generating the adjacency graph, we can use a complete graph of N vertices since each pair of news stories can be related. But this may cause problems since a complete graph of N vertices has $\binom{N}{2} = O(N^2)$ edges and the number of all possible solutions is exponential in the number of edges, i.e. $O(2^{N^2})$, which requires a long convergence time. By investigating the data, however, one may observe that some story pairs have few similarities in terms of contents. Such pair of stories shall never be grouped together. So graph pruning can be performed on the adjacency graph before SWC . We define a threshold τ, and cut all edges e whose $\mathcal{D}(e) \geq \tau$ deterministically.

5 Topic Tracking

In this section, we describe our method for tracking topics detected in certain continuous time periods. We link all detected topics in different time periods to form topic trajectories over time.

We divide the whole news data collection into several sub-collections which consist of news stories in different time periods. Topic detection is performed within each sub-collection separately. The sub-collection set of the news corpus D is denoted by $\{C_1, ..., C_M\}$ where $C_1 \cup ... \cup C_M = D$ and M is the number of sub-collections. Each sub-collection contains news documents from one specific time span t_i. Topics extracted within each sub-collection C_i are denoted by $\mathbf{\Theta_i} = \{\Theta_i^1, ..., \Theta_i^{K_i}\}$, where K_i is the obtained topic number.

For topic tracking, we link topics detected in the sub-collections. One optional method for solving the linking problem is to do another clustering on the detected topics using SWC. But to fast obtain the topic links, we choose to measure the similarities between topics by considering both the topic content similarities and their temporal distances, and use a threshold to decide whether they can be linked. Formally, the similarity measurement to decide whether two topics can be linked is calculated as:

$$\begin{aligned} Sim(\Theta_{i_1}^{k_1}, \Theta_{i_2}^{k_2}) = \alpha_{sim} \exp\{-\beta_{kl}[KL(\Theta_{i_1}^{k_1} || \Theta_{i_2}^{k_2}) \\ + KL(\Theta_{i_2}^{k_2} || \Theta_{i_1}^{k_1})]\} + (1 - \alpha_{sim}) \exp\{-|t_{i_1} - t_{i_2}|\}, \end{aligned} \tag{19}$$

where $i_1 \neq i_2$, and α_{sim} and β_{kl} are positive parameters. Note that using the proposed topic representation, each topic is composed of the image part and the text part, and they can be further divided into the "who", "where" and "what" components, and the face and object components respectively. Thus we have five components in total. Each component is represented using one model. The KL divergence of one topic given another is therefore averaged over these models:

$$KL(\Theta_{i_1}^{k_1} || \Theta_{i_2}^{k_2}) = \sum_{j=1}^{5} \lambda_j KL(\Theta_{m_1}^{k_1,j} || \Theta_{i_2}^{k_2,j}), \tag{20}$$

where λ_j is the corresponding weight, and $\Theta_{i_1}^{k_1,j}$, $\Theta_{m_2}^{k_2,j}$ are the histograms of word frequencies for the j-th component. After calculating the topic similarities using Eq. 19, a threshold τ_{link} can be used for pruning the links between topics to get the final topic trajectories.

6 Experiments

6.1 Datasets

Two datasets are used in our experiment:

1) **Reuters-21578.** Reuters-21578 dataset[4] is a publicly available collection of news stories from Reuters newswire. It is widely used for the evaluation of clustering and classification methods. The dataset contains 21,758 stories which belong to 135 clusters/categories. The clusters/categories are annotated manually. Only textual information is contained in the dataset.

2) **UCLA Broadcast News Dataset.** We collected a multimedia broadcast news dataset from UCLA Library Broadcast NewsScape. Five US networks are included in the dataset: CNN, MSNBC, FOX, ABC, and CBS. It contains 379 news videos broadcasted in the time period from June 1, 2013 to June 14, 2013. The total length of the videos is about 362 hours. Several programs from each news network are included in the dataset, such as "CNN Newsroom", "MSNBC News Live", "FOX Morning News", "ABC Nightline", "CBS News", etc.

Annotation: We annotate the UCLA Broadcast News Dataset for topic detection and tracking. We let annotators decide whether a pair of stories belong to the same topic or not. 10,000 story pairs are annotated by three

4. Reuters-21578 dataset can be downloaded at http://www.daviddlewis.com/resources/testcollections/reuters21578/.

annotators. For each story pair we treat the relation that most annotators agree as the ground-truth relation[5].

This dataset is mainly used for quantitative evaluation. To show how our method work on large-scale datasets qualitatively, we also apply it to more news data from the UCLA Library Broadcast NewsScape.

6.2 Experiment I: Topic Detection

In this experiment, we conduct topic detection experiments on both the Reuters-21578 dataset and the UCLA Broadcast News Dataset.

6.2.1 Results on Reuters-21578

We compare the proposed topic detection method with other story/document clustering methods on the news dataset Reuters-21578 (only texts are available). Stories with multiple cluster labels are discarded and for the remaining stories, only those from the largest 10 clusters are selected [12].

Evaluation Protocol. On Reuters-21578, we follow the evaluation protocol in [12], [64]. Two metrics are used to evaluate the clustering performance, i.e. accuracy, and normalized mutual information. To compute the accuracy, the obtained clusters are mapped to the ground-truth clusters in the dataset. The clustering accuracy is then defined as the percentage of documents that have the correct cluster labels after mapping. The mutual information measures the mutual dependence of the ground-truth cluster assignments and the obtained clustering assignments of the documents. Please refer to [64] for more detailed definitions of these two metrics.

Other Methods. Several other methods are included in the comparison, namely: (1) K-means and Normalized Cuts (NC) [65], which are widely used clustering and graph partitioning algorithms; (2) Nonnegative-Matrix-Factorization (NMF) based clustering [66], Latent Semantic Indexing (LSI) [18], and Locally Consistent Concept Factorization (LCCF) [64], which are factorization based methods that are effective for document clustering; (3) LDA + K-means [12] and LDA + Naive [12] (both of them use LDA to learn topics and topic distributions for documents, and LDA + K-means then clusters documents using K-Means based on these distributions while LDA + Naive treats the label of the most dominant topic as the cluster label for each document); and (4) Multi-grain clustering topic model (MGCTM) [12] which has the best clustering result on Reuters-21578 so far. The inputs of these methods in the comparison are the documents' tf-idf vectors [12], [64]. These methods all require the cluster number to be specified in the input. Thus for these methods, we set the cluster number $K = 10$ in the experiment, which equals the ground-truth cluster number in the dataset. Please refer to [12] for other detailed settings of these algorithms.

5. We illustrate why we choose to annotate story pairs and how the 10,000 story pairs are chosen in the supplementary material.

Table 1: Clustering Performance of different methods on Reuters-21578.

	Clustering Accuracy(%)	Normalized Mutual Information(%)
K-Means	35.02	35.76
NC [65]	26.22	27.40
NMF [66]	49.85	35.89
LSI [18]	42.00	37.14
LCCF [64]	33.07	30.45
LDA + K-means [12]	29.73	36.00
LDA + Naive [12]	54.88	48.00
MGCTM [12]	56.01	50.10
Our method	**67.19**	**51.97**

Parameter Settings of Our Method. To compare with the other algorithms, in our method, we add a Gaussian prior term with the mean $\mu = 10$ and variance $\sigma^2 = 0.5$ to Eq. 13 to make the sampling process converge to the state where the cluster number equals 10. The parameter α in Eq. 13 is set as $\alpha = 0.2$. The weights $\{\lambda_F, F \in \mathcal{F}\}$ in Eq. 15 are set as: $\lambda_{F_{who}} = 0.1$, $\lambda_{F_{where}} = 0.1$, $\lambda_{F_{what}} = 0.4$, $\lambda_{F_{face}} = 0.1$ and $\lambda_{F_{object}} = 0.3$. The threshold τ used for graph pruning is set as $\tau = 160$.

Comparison Results. Table 1 shows the results of different methods on Reuters-21578. It can be seen from the results that our approach is better than other methods in terms of both the clustering accuracy and the normalized mutual information. This is because our method uses the MT-AOG representation which organizes topics in a hierarchical way and embeds contexts between different components. The cluster sampling method SWC also plays an important role in optimizing the solution. Other methods generally use the basic word distributions and most of the solutions they get are locally optimal.

6.2.2 Results on UCLA Broadcast News Dataset

We conduct both qualitative and quantitative evaluations of our topic detection method on the UCLA Broadcast News Dataset. We preprocess news videos and closed captions to obtain texts and key frames used in the experiment[6]. After preprocessing, we have 3,633 news stories including 577,721 words and 36,810 keyframes. The whole collection contains 24,036 unique word terms.

1) Qualitative Evaluation. We conduct the topic detection experiment on the whole dataset.

Parameter Settings. The parameter α in Eq. 13 is set as $\alpha = 10$. The "fast" mode of Selective Search is used to generate the possible object patches (please refer to [59] for more detailed settings of the "fast" mode). The cluster numbers for grouping the faces and object patches in Section 3 are set as $1,000$ and $1,500$ respectively. We also delete clusters with a small number

6. The preprocessing steps are illustrated in the supplementary material. Preprocessing visual frames such as face detection or running a convolutional neural network is in fact more time-consuming part in our system. It takes about 5s to preprocess one visual frame. In practice, we use a distributed computing system to extract the features.

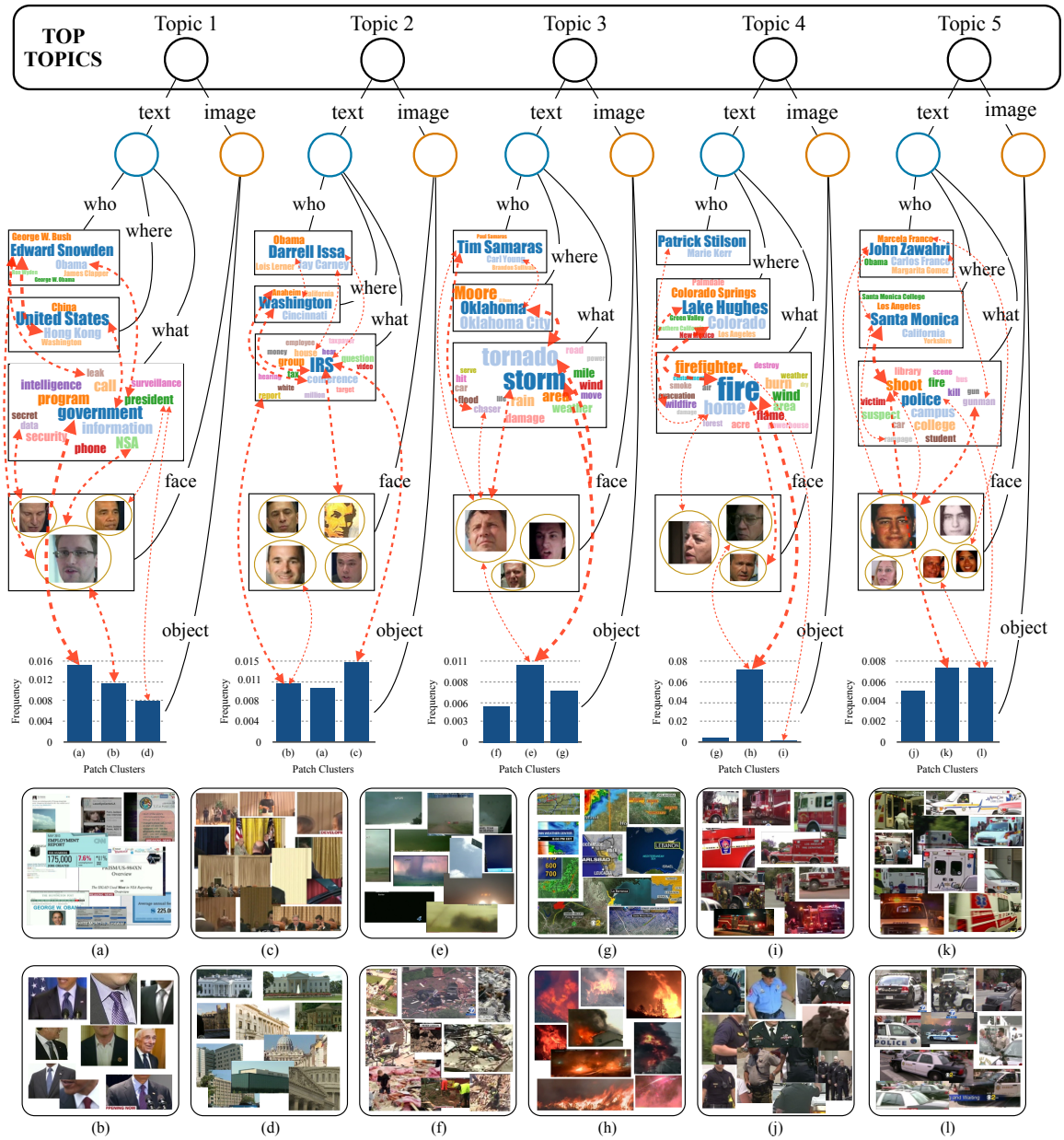


Figure 7: Top five topics detected in the dataset we collected. The top words for who, where and what components are shown with their sizes proportional to their frequencies. The top faces and objects are also shown. The face sizes are also proportional to their frequencies. Object clusters are shown in squares at the bottom part of the figure. The objects' frequencies in the topic are shown by the curves above the squares. The dashed red lines show the top co-occurring pairs between different components and the thickness of each line is proportional to the related pair frequency.

of patches. The remaining cluster numbers for face and object are 708 and $1,316$ respectively. Other parameter settings are the same as those in Section 6.2.1.

Topic Detection Results. We show the detected top five topics in Fig. 7. Topic 1 talks about the news that Edward Snowden leaked information from National Security Agency (NSA). Topic 2 is about the IRS scandal, including the discussion on the misuse of taxpayers' money and the related hearing. Topic 3 mainly talks about the Oklahoma tornado, including its development, the damage it caused, and the storm chasers' stories. Topic 4 is about the wildfires, which also includes the fire development and the related damages. Topic 5 is about the Santa Monica College shooting rampage, and the related gunman and victims' stories are also included. We can see from the figure that the obtained structured results can clearly describe the related topics. The involved people's names and face patches, related locations, key objects, descriptions about the event, as well as the co-occurrence relations between them (represented by the dashed red lines) are all shown in the structure. And as shown in Fig. 7, the topic components and their co-

occurrence relations are closely related to the detected topic.

2) Quantitative Evaluation.

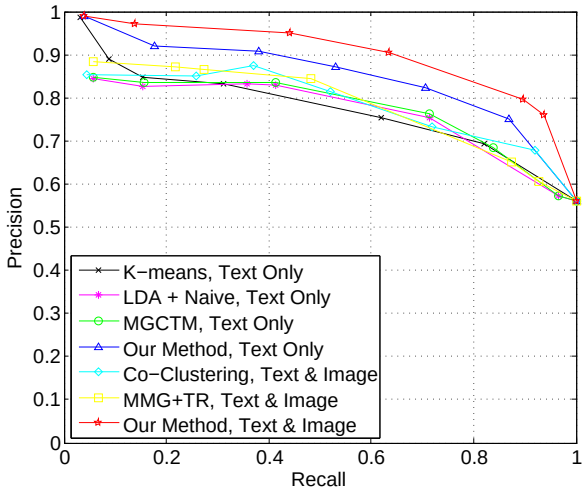
Evaluation Protocol. Using the annotated story pairs, we draw precision-recall curves for different topic detection methods in the evaluation. The precision is calculated as the fraction of story pairs that actually belong to one topic out of those that are computed to be. The recall is the fraction of story pairs that are computed to belong to one topic out of those that actually do.

Other Methods. Among the comparison methods used in 6.2.1, we select two methods with better performance, i.e. LDA + Naive and MGCTM. We also include the widely used k-means algorithm. These algorithms are all unimodal, so their inputs in the experiment are the stories' textual information, i.e. the stories' tf-idf vectors. Two multimodal methods are also included in the comparison, including the multimodal co-clustering method in [14], and the multimodality graph with topic recovery method (MMG+TR) in [38]. For these method, we set a sequence of cluster numbers in the experiment to generate the precision-recall curves.

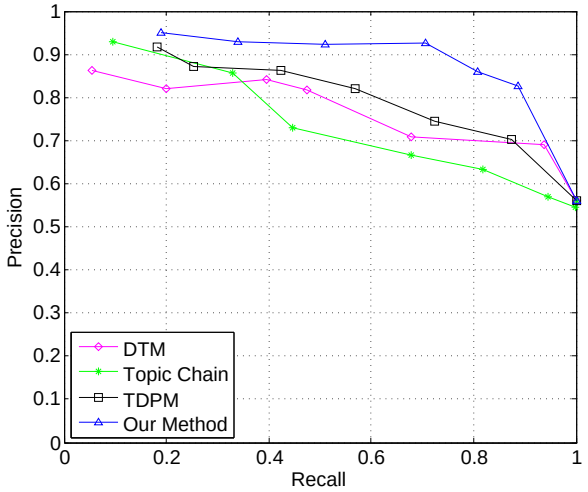
Parameter Settings of our method. To generate the precision-recall curve, we vary the parameter α_K in Eq. 13 for our method. Other parameter settings are the same as those in the qualitative experiment. To compare with the unimodal methods, we also conduct experiments where only the text information is included.

Comparison Results. Fig. 8a shows the precision-recall curves for different methods. As we can see from the figure, based on merely text information, our method has better performance than the other unimodal methods. This shows that the proposed Multimodal Topic And-Or Graph (MT-AOG) and the clustering sampling method we use can help generate better topics. The comparison results of k-means, LDA + Naive, and MGCTM are similar to those in Reuters-21578. With the visual information added, our performance gets further improved, showing the effectiveness of our method which jointly models texts and images. Multimodal methods also perform better than the other unimodal methods in general, which shows the necessity of using visuals in topic detection.

Evaluation of Contextual Relations. To demonstrate the effectiveness of contextual relations in MT-AOG, we conducted an ablation study. The contextual relations in the text part (Sec. 3.2), in the image part (Sec. 3.3), and between these two parts (Sec. 3.4) were tested in the experiment. Note that there are 303 stories in our dataset (8.34%) which do not have any image elements (e.g., only anchor's comments without field footage). These stories were treated as individual clusters in "image-only" cases. The results are shown in Fig. 9. As expected, incorporating contextual relations is critical for achieving a better clustering performance in all cases, which justifies our model choice. In addition, this result also reinforces that using multimodal cues together is better than using a single channel.



(a) Topic Detection.



(b) Topic Tracking.

Figure 8: Precision-recall curves of our topic detection and tracking method and comparisons with other methods on UCLA Broadcast News Dataset.

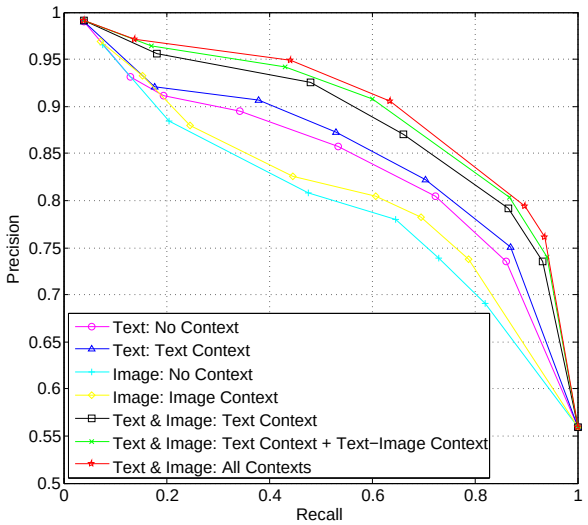


Figure 9: Precision-recall curves of our topic detection methods with/without different contextual relations.

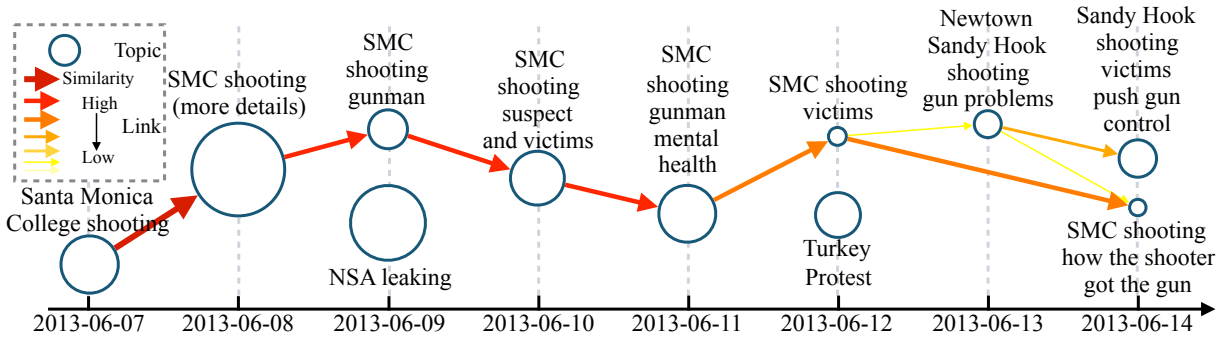


Figure 10: Topic tracking result of the event Santa Monica Shooting. Each circle represents one topic and the circle size is proportional to the size of the topic, i.e. the volume of corresponding news stories. Thicker links represent greater similarities between topics.

6.3 Experiment II: Topic Tracking

In this experiment, we conduct topic tracking experiments on the UCLA Broadcast News Dataset. Both qualitative and quantitative evaluations of our method are included in the experiment.

1) Qualitative Evaluation. To show that our topic tracking method can generate meaningful topic trajectories, we conduct the qualitative evaluation experiment.

Parameter Settings. To track topics over time, we divide the dataset into 14 sub-collections each of which contains news stories from one day. The average number of news stories per day is 260, and these stories on average contain 41,266 words and 2,629 keyframes. Topic detection is firstly performed within each sub-collection. Then given the detected topics, we do topic tracking, which links topics over time and generates topic trajectories. The parameter α_{sim} and β_{kl} in Eq. 19 are set as $\alpha_{sim} = 0.8$ and $\beta_{kl} = 0.005$ respectively. The weights $\{\lambda_i; i = 1, ..., 5\}$ in Eq. 20 are set as $\{0.1, 0.1, 0.4, 0.1, 0.3\}$. The threshold τ_{link} for selecting links between topics is set as $\tau_{link} = 0.7$.

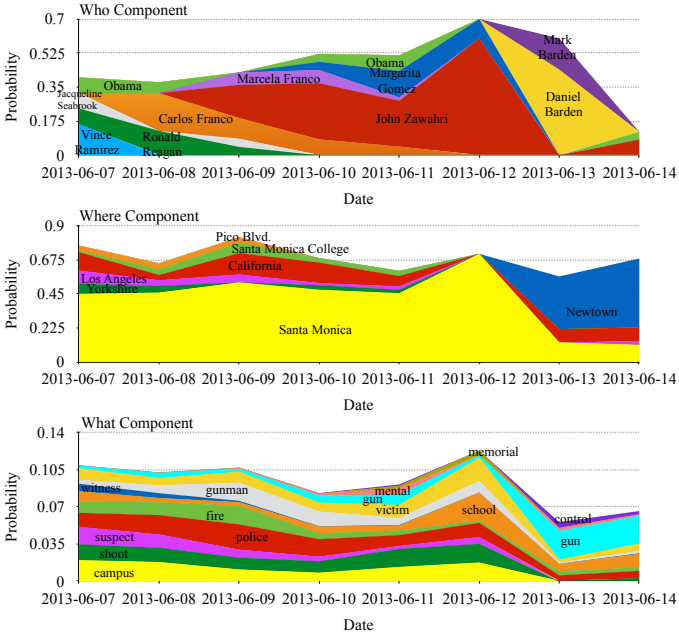
Topic Tracking Results. One topic tracking trajectory about the Santa Monica College shooting is shown in Fig. 10. The topics are summarized in several words here for space constraints. The descriptions of the text part and the image part for the corresponding topics in the trajectory are shown in Fig. 11a and Fig. 11b respectively. The probabilities of the top textual and visual words over time are shown in the figure.

Our experiment is conducted on a computer with 3.6 GHz CPU and 16G RAM. The average time for topic detection for one day's stories is 7.16s, and the average time for topic tracking is 2.41s. So our method can deal with news streams efficiently.

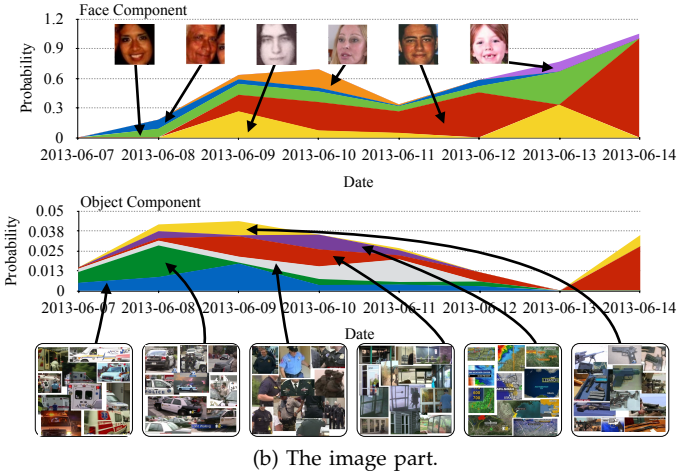
2) Quantitative Evaluation. We also conduct quantitative evaluation on our topic tracking method.

Evaluation Protocol. For topic tracking, we also use the precision-recall curves to compare different methods.

Other Methods. We include three methods in the comparison, namely: (1) Dynamic topic model (DTM) [21] which models topic changes over time; (2) topic chain method [43] which generates topics in different



(a) The text part.



(b) The image part.

Figure 11: The text and image parts of the topics corresponding to the trajectory shown in Fig. 10. The probabilities of top words and faces/objects along the time span are shown.

time periods using LDA and links these topics to form topic chains; and (3) temporal Dirichlet mixture model (TDPM) [41] for evolutionary clustering. These three methods are all unimodal. For DTM, we set different topic numbers to generate its precision-recall curve. For the topic chain method, we set the topic number in each time period as 50 and use a sequence of similarity threshold when building the topic chains. For TDPM, we vary the concentration parameter to obtain its precision-recall curve.

Parameter Settings of our method. We vary the parameter τ_{link} to generate the precision-recall curve.

Comparison Results. Fig. 8b shows the precision-recall curves for our tracking method and the other two methods. Our method outperforms the other two methods since both texts and images are included. Moreover, our topic detection method can generate meaningful topics, which is also an important factor for the topic tracking performance.

6.4 Experiment III: Large-Scale Topic Detection and Tracking

To show that our method can work effectively on large-scale datasets qualitatively, we conduct topic detection and tracking experiment using the CNN news data in 2012. We firstly detect topics within each month, and then do topic tracking among the detected topics.

The obtained trajectories are shown in Fig. 12. Due to the space limit we only show text parts of topics in the trajectory. The top part of the figure shows the trajectory of George Zimmerman's case and some other related shooting cases such as the Chardon shooting in February and the Colorado theater shooting in July. The middle part of the figure is mainly about topics closely related to the 2012 US election, such as the health care, the immigration problem, the economy and the debates. The Syria problem, which is another factor related to the election, is shown in the bottom part of the figure. We also get some short trajectories such as the one about Olympic shown in the lower half part of the figure. The "who", "where", and "what" components shown in the figure are highly relevant to the corresponding topics. From the topic trajectories, we can clearly see how these topics develop over time and how they relate to each other. For instance, from the trajectory of George Zimmerman's case, we can see that the shooting happened in Feb. 2012, followed by the bond hearing and defense process, and the final trial in July 2012. This case is related to other shooting cases shown by the links between them.

We also made a case study which tracks the 2016 U.S. presidential election. The large-scale topic detection and tracking results are visualized in the Viz2016 website (mentioned in Section 1.1).

7 Conclusions

We have presented a joint image-text news topic detection and tracking method. We propose a structured topic representation, i.e. Multimodal Topic And-Or Graph, which models image and text parts of multiple topics jointly. We detect topics using the SWC-based cluster sampling method. Topics are also tracked over time to deal with continuous updates of news streams. Both qualitative and quantitative evaluation results show the effectiveness and efficiency of the proposed method over existing methods.

In the future, we will expand our study to concrete media analysis for social and political science research. Based on our topic detection and tracking results, we can analyze, for example, how media outlets are biased for different topics, what is the agenda-setting pattern, what is the causal relations between topics, etc.

Acknowledgment

This project is supported by the NSF CDI project CNS 1028381. The authors would like to thank Dr. Francis Steen and Tim Groeling at UCLA, and Dr. Chengxiang Zhai at UIUC for discussions and insightful suggestions. We would also like to thank Dr. Quanshi Zhang and Tianfu Wu at UCLA for their assistance.

References

[1] J. Allan, *Topic Detection and Tracking: Event-based Information Organization*. Norwell, MA, USA: Kluwer Academic Publishers, 2002.
[2] J. Joo, W. Li, F. Steen, and S.-C. Zhu, "Visual persuasion: Inferring communicative intents of images," in *CVPR*, 2014, pp. 216–223.
[3] J. Joo, F. Steen, and S.-C. Zhu, "Automated facial trait judgment and election outcome prediction: Social dimensions of face," in *ICCV*, 2015, pp. 3712–3720.
[4] S.-C. Zhu and D. Mumford, "A stochastic grammar of images," *Found. Trends. Comput. Graph. Vis.*, vol. 2, no. 4, pp. 259–362, 2006.
[5] A. Barbu and S.-C. Zhu, "Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities," *TPAMI*, vol. 27, no. 8, pp. 1239–1253, 2005.
[6] J. Joo, S. Wang, and S.-C. Zhu, "Hierarchical organization by And-Or Tree," in *Book chapter in Handbook of Perceptual Organization*, J. Wagemans, Ed. Springer, 2012.
[7] B. Z. Yao, B. X. Nie, Z. Liu, and S.-C. Zhu, "Animated pose templates for modeling and detecting human actions," *TPAMI*, vol. 36, no. 3, pp. 436–452, 2014.
[8] G. Hart, "The five W's: An old tool for the new task of task analysis," *Technical Communication*, vol. 43, no. 2, pp. 139–145, 1996.
[9] J. Harriss, K. Leiter, and S. P. Johnson, *The Complete Reporter: Fundamentals of News Gathering, Writing, and Editing, Complete with Exercises*. MacMillan Publishing Company, 1981.
[10] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report," in *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 194–218.
[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[12] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling," in *UAI*, 2013, pp. 694–703.
[13] Y. Zhai and M. Shah, "Tracking news stories across different sources," in *MM*, 2005, pp. 2–10.
[14] X. Wu, C.-W. Ngo, and Q. Li, "Threading and autodocumenting news videos: a promising solution to rapidly browse news topics," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 59–68, 2006.
[15] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating topics and syntax," in *NIPS*, 2005.
[16] J. Boyd-graber and D. Blei, "Syntactic topic models," in *NIPS*, 2009.

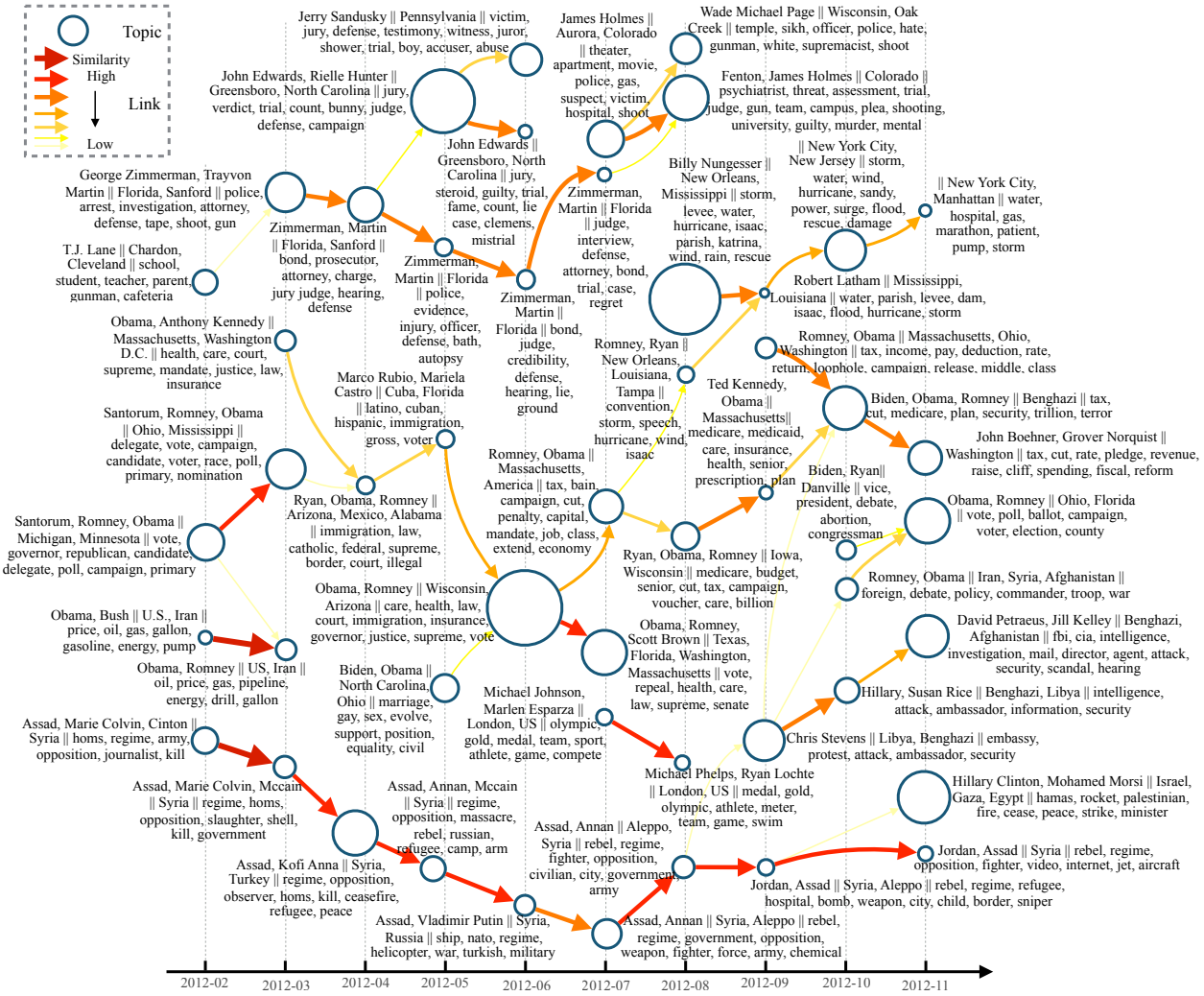


Figure 12: Topic trajectories for 2012 CNN news. Each circle represents one detected topic and the circle size is proportional to the topic size, i.e. the volume of corresponding news stories. Thicker links correspond to greater similarities between topics. The who, where and what parts of the topic are separated by the symbol "||".

[17] X. Wu, C.-W. Ngo, and A. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *TMM*, vol. 10, no. 2, pp. 188–199, 2008.

[18] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999, pp. 50–57.

[19] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[20] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, pp. 5228–5235, 2004.

[21] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*, 2006.

[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *JASA*, vol. 101, no. 476, pp. 1566–1581, 2006.

[23] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *SIGKDD*, 2011, pp. 448–456.

[24] D. Newman, C. Chemudugunta, and P. Smyth, "Statistical entity-topic models," in *SIGKDD*, 2006, pp. 680–686.

[25] Z. Li, B. Wang, M. Li, and W.-Y. Ma, "A probabilistic model for retrospective news event detection," in *SIGIR*, 2005, pp. 106–113.

[26] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR*, 2003, pp. 127–134.

[27] D. Putthividhy, H. Attias, and S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *CVPR*, 2010, pp. 3408–3415.

[28] Y. Zhou and J. Luo, "Geo-location inference on news articles via multimodal plsa," in *MM*, 2012, pp. 741–744.

[29] H. Cai, Y. Yang, X. Li, and Z. Huang, "What are popular: Exploring twitter features for event detection, tracking and visualization," in *MM*, 2015, pp. 89–98.

[30] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *TMM*, vol. 18, no. 2, pp. 233–246, 2016.

[31] T. Chen, H. M. SalahEldeen, X. He, M.-Y. Kan, and D. Lu, "VELDA: Relating an image tweet's text and images," in *AAAI*, 2015, pp. 30–36.

[32] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *ICCV*, 2011, pp. 2407–2414.

[33] J. Chang and D. M. Blei, "Relational topic models for document networks," in *AIStats*, vol. 9, 2009, pp. 81–88.

[34] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Semi-supervised relational topic model for weakly annotated image recognition in social media," in *CVPR*, 2014, pp. 4233–4240.

[35] J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. V. Mulbregt, "Topic tracking in a news stream," in *In Proceedings of DARPA Broadcast News Workshop*, 1999, pp. 133–136.

[36] C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining Text Data*. Springer US, 2012, pp. 77–128.

[37] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *In KDD Workshop on Text Mining*, 2000.

[38] L. Chu, Y. Zhang, G. Li, S. Wang, W. Zhang, and Q. Huang, "Effective multi-modality fusion framework for cross-media topic

detection," *TCSVT*, vol. PP, no. 99, pp. 1–1, 2014.
[39] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," *Inf. Retr.*, vol. 7, no. 3-4, pp. 347–368, 2004.
[40] W. Hsu and S.-F. Chang, "Topic tracking across broadcast news videos with visual duplicates and semantic concepts," in *ICIP*, 2006.
[41] A. Ahmed and E. P. Xing, "Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering." in *SDM*, 2008, pp. 219–230.
[42] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *SIGKDD*, 2005.
[43] D. Kim and A. Oh, "Topic chains for understanding a news corpus," in *International Conference on Computational Linguistics and Intelligent Text Processing*, 2011.
[44] B. Jou, H. Li, J. G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang, "Structured exploration of who, what, when, and where in heterogeneous multimedia news sources," in *MM*, 2013.
[45] H. Li, B. Jou, J. G. Ellis, D. Morozoff, and S.-F. Chang, "News rover: Exploring topical structures and serendipity in heterogeneous multimedia news," in *MM*, 2013.
[46] M. C. Yu, P. Vajda, D. M. Chen, S. S. Tsai, M. Daneshi, A. F. Araujo, H. Chen, and B. Girod, "Eigennews: A personalized news video delivery platform," in *MM*, 2013.
[47] M. Daneshi, P. Vajda, D. Chen, S. Tsai, M. Yu, A. Araujo, H. Chen, and B. Girod, "Eigennews: Generating and delivering personalized news video," in *ICMEW*, 2013.
[48] C. A. Bejan and S. Harabagiu, "Unsupervised event coreference resolution," *Computational Linguistics*, vol. 40, no. 2, pp. 311–347, 2014.
[49] T. Zhang, H. Li, H. Ji, and S.-F. Chang, "Cross-document event coreference resolution based on cross-media features," in *EMNLP*, 2015.
[50] H. Ji and R. Grishman, "Knowledge base population: Successful approaches and challenges," in *ACL*, 2011, pp. 1148–1158.
[51] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the TAC 2010 knowledge base population track," in *TAC*, 2010.
[52] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. ACM Press, 2006, pp. 321–330.
[53] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition." in *LREC*, 2012.
[54] A. Araujo, J. Chaves, D. Chen, R. Angst, and B. Girod, "Stanford i2v: a news video dataset for query-by-image experiments," in *MMSys*, 2015.
[55] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *ACL*, 2005, pp. 363–370.
[56] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.
[57] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *TPAMI*, vol. 28, no. 12, pp. 2037–2041, 2006.
[58] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition," in *ICCV*, vol. 1, 2005, pp. 786–791.
[59] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.
[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012.
[61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, pp. 1–42, 2015.
[62] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *MM*, 2014, pp. 675–678.
[63] M. Pavlovskaia, K. Tu, and S.-C. Zhu, "Mapping the energy landscape of non-convex optimization problems," in *EMMCVPR*, 2015, pp. 421–435.
[64] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *TKDE*, vol. 23, no. 6, pp. 902–913, 2011.
[65] J. Shi and J. Malik, "Normalized cuts and image segmentation," *TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.
[66] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *SIGIR*, 2003, pp. 267–273.

Weixin Li is currently pursuing the Ph.D. degree in Computer Science with the Center for Vision, Cognition, Learning and Autonomy (VCLA) at the University of California, Los Angeles (UCLA). She received the MS degree in Computer Science from UCLA in 2014, and the B.E. degree in Computer Science from Beihang University in 2010. Her current research interests include computer vision and big data analytics, with a focus on the study of news and social media.

Jungseock Joo received the BSE degree in computer science and engineering from Seoul National University in 2008, the MS degree from Columbia University in 2010, and the PhD degree in computer science from UCLA in 2015. He was a research scientist at Facebook in 2015-2016. He is now an assistant professor in the Department of Communication Studies at UCLA. His research interests include computer vision, multimodal communication, and computational media analysis. He was a recipient of International Kwanjeong Fellowship while he was a PhD student at UCLA in 2010-2015.

Hang Qi is currently a Ph.D. student in Computer Science and graduate student researcher at Center for Vision, Cognition, Learning and Autonomy (VCLA) at the University of California, Los Angeles (UCLA). His research interests include computer vision, artificial intelligence, and machine learning. He received a MS degree in Computer Science from UCLA in 2013 and BEng degree in Software Engineering from Tongji University in 2012.

Song-Chun Zhu received a Ph.D. degree from Harvard University in 1996. He is currently a professor of Statistics and Computer Science at UCLA, and the director of the Center for Vision, Cognition, Learning and Autonomy. He has published over 160 papers in computer vision, statistical modeling and learning, cognition, and visual arts. He received a number of honors, including the J.K. Aggarwal prize from the Int'l Association of Pattern Recognition in 2008 for "contributions to a unified foundation for visual pattern conceptualization, modeling, learning, and inference", the David Marr Prize in 2003 with Z. Tu et al. for image parsing, twice Marr Prize honorary nominations in 1999 for texture modeling and in 2007 for object modeling with Z. Si and Y.N. Wu. He received the Sloan Fellowship in 2001, a US NSF Career Award in 2001, and an US ONR Young Investigator Award in 2001. He received the Helmholtz Test-of-time award in ICCV 2013, and he is a Fellow of IEEE since 2011.