International Journal of Research
Available at
https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

Graph Approach Markov Assumptions for Social LDA Inspection

V. LAKSHMI SWATHI, G. SUBBA RAO
PG Student, Dept. of MCA, St. Ann's college of Engineering &Technology , chirala.
Associate Prof, Dept. of MCA, St. Ann's college of Engineering &Technology , chirala.

Abstract: *Content analysis is applied social science research method is increasingly being supplemented by topic modeling. This approach 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. We find news topics through a cluster sampling process which groups stories about closely related events together. Swenson Wang Cuts (SWC), an effective cluster sampling algorithm, Our system demonstrate that incorporating event information in the prediction tasks reduces the root mean square error (RMSE) of prediction by 22% compared to the standard ARIMA model. We present a method for automatically collecting television news and social media content (Twitter) and discovering the hash tags that are relevant for a TV news video. Our algorithms incorporate both the visual and text information within social media and television content and improve performance over single modality methods. This paper present LDA-style topic model that captures not only the low-dimensional structure of data, structure changes over time Unlike other recent work that relies on Markov assumptions of time here each topic is associated with a continuous distribution over timestamps, and for each generated document the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp.*

Index Terms: Graphical Models, Temporal Analysis, Topic Modeling, ash tagging; social media, Multimodal Topic And-Or Graph, cluster sampling.

1. INTRODUCTION

NEWS stories provide data in 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 [1]. The primary objective of this paper is to develop automatic topic detection and tracking method which can be used to analyze the real world events and their relationships. 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. These events dynamically introduce new people or new places involved and are eventually connected to other events.

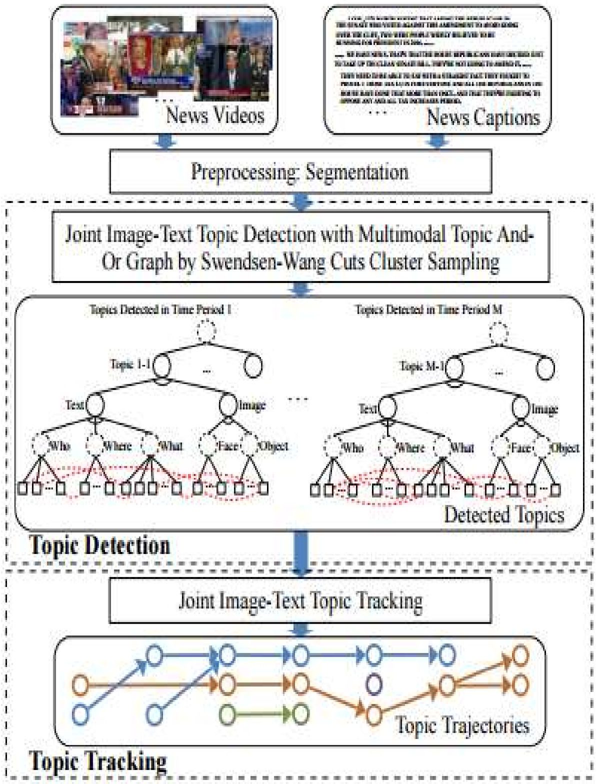

News Videos
News Captions
Preprocessing: Segmentation
Joint Image-Text Topic Detection with Multimodal Topic And-Or Graph by Swendsen-Wang Cuts Cluster Sampling
Topics Detected in Time Period 1
Topics Detected in Time Period M
Topic 1-1
Topic M-1
Text
Image
Who
Where
What
Face
Object
Detected Topics
Topic Detection
Joint Image-Text Topic Tracking
Topic Trajectories
Topic Tracking

Figure 1: Overview detected topics are tracked over trajectories

This paper presents Topics over Time (TOT), a topic model that explicitly models time jointly with word co occurrence patterns. Significantly, and unlike some recent work with similar goals, our model does not discredited time, and does not make Markov assumptions over state transitions in time. Rather, TOT parameterizes a continuous distribution over time associated with each topic [4], and topics are responsible for generating both observed timestamps as well as words. Parameter estimation is thus driven to discover topics that simultaneously capture word co-occurrences and locality of those patterns in time. When a strong word co-occurrence pattern appears for a brief moment in time then disappears, TOT will create a topic with a narrow time distribution [5]. Twitter users link their tweets to a subject by using hash tags. A hash tag, is a short word or acronym preceded by the "#" sign. Hash tags can then be searched easily within social media sites, making it simple to quickly obtain cogent information about a possibly very narrow topic or opinion in an efficient manner [6]. Although social media platforms have become major sources of news, especially for young people, traditional TV news still provides a unique value to customers. Twitter attracts attentions from major news agencies and TV channels as it can provide an increased reach for their content. However, the content from TV broadcasts has to be properly integrated in the social media context in order to maximize this reach. By linking the broadcast news videos to hash tags [7].

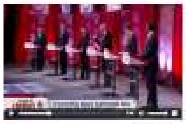
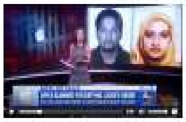
Frame from video	Found hashtags
	#politics, #trump2016, #realdonaldtrump, #goptownhall, #votetrumpsc
	#politics, #dumptrump, #realdonaldtrump, #gopdebate
	#iphone, #apple, #cybersecurity

Fig No 2. Hash tags in red were found predominantly using our text pipeline, and those in blue were found using predominantly visual pipelines

2. RELATED WORK

Several previous studies have examined topics and their changes across time. Rather than jointly modeling word co occurrence and time, many of these methods use post-hoc or pre-discredited analysis. The first style of non-joint modeling involves fitting a time-unaware topic model, and then ordering the documents in time, slicing them into discrete subsets, and examining the topic distributions in each time-slice. One example is Griffiths and Severs' study of PNAS proceedings [8], in which they identified hot and cold topics based on examination of topic mixtures estimated from an LDA model. The second style of non-joint modeling pre-divides the data into discrete time slices, and fits a separate topic model in each slice. Examples of this type include the experiments with the Group-Topic model [9], in which several decades. An interesting and challenging problem using Twitter data is event detection and tracking. Topic Detection and Tracking (TDT) is a traditional research topic, which focus on finding events and linking documents to the related event in broadcast news streams [10]. However, TDT on twitter data faces new challenges due to the limited number of characters used in twitter messages and also the large number of noise and meaningless messages on twitter [11]. Several advanced methods have been developed recently to address the event detection and tracking problem on twitter data, to detect unspecified types of events discover predefined types of events from twitter messages. Hash tags are used by social media users to convey an idea we use hash tags to organize tweets as topics and link them to broadcasting TV news events [12]. Our work differs from the aforementioned in that we propose a framework, where we specifically connect real-world events extracted from text data (news articles) to predict external variables or indicators, driven by the assumption that there exists a dependency between these variables and real-world events. In a sense, our framework is similar to Google Correlate [13], in terms of general applicability, however we focus on identifying real-world events in news sources whereas Google Correlate focuses on web search logs to determine queries that are correlated with real-world phenomenon. In addition, we propose a novel generative model to identify events that drive fluctuations in socio-economic indicators, whereas Google Correlate uses standard correlation metrics.

3. EVENT CLASS MODEL

For any generic document, ranging from academic articles to blog posts, topics can be a very important and useful feature for understanding its content. Specialized documents like news articles whose purpose is to report stories and incidents, we claim that events, which are aimed at capturing the main "actions" is more informative than topics, which essentially capture the main themes in the document [14]. We envision topics as being part of the event description but there are additional aspects of events that need to be captured separately. An event is an amalgamation of many components entities, topics and metadata like location and time. We focus on modeling only the underlying essence of any event the action words that are representative of incidents reported in the article. This is in contrast to topic models that consider all the words mentioned in a document [15].

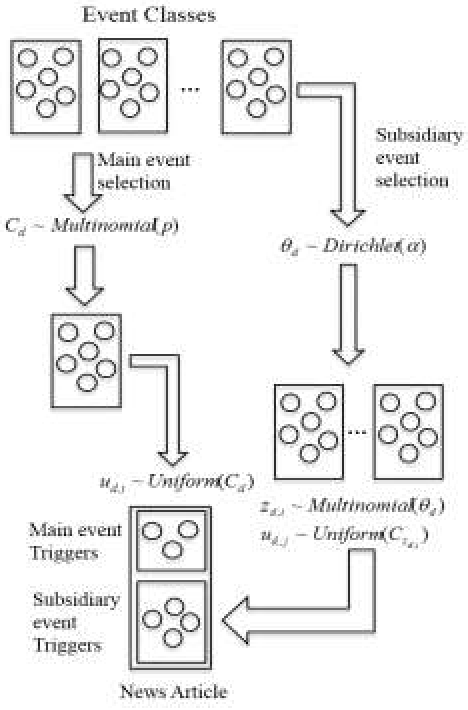

Event Classes
...
Main event selection
Subsidiary event selection
$C_d \sim Multinomial(p)$
$\theta_d \sim Dirichlet(\alpha)$
$u_{d,i} \sim Uniform(C_d)$
...
$z_{d,i} \sim Multinomial(\theta_d)$
$u_{d,j} \sim Uniform(C_{z_{d,i}})$
Main event Triggers
Subsidiary event Triggers
News Article

Figure 3: Event class model and generation of news articles.

For each article d ∈ D, its event class Cd is sampled from a categorical distribution with parameter p: Cd ∼ Categorical(p) where p = (p1, p2, . . . , pK) denotes the prior probability distribution of the event classes in the corpus D. 2. For each event class trigger udi ∈ U e d in the article d, sample an event trigger uniformly at random from the triggers belonging to sampled event class: udi ∼ Uniform(Cd) where Uniform(Cd)is the uniform distribution over the event triggers in the set Cd ∈ C. 3. After the main event class along with the main event triggers are generated, the subsidiary events are sampled. The subsidiary events generation is similar to topic generation in LDA.

4. TOPICS OVER TIME

The Topics over Time (TOT) model review based on Latent Dirichlet Allocation model. Our notation is summarized in the graphical model representations of both LDA and our TOT models. In TOT model to discovery is influenced not only by word co-occurrences, but also temporal information. Rather than modeling a sequence of state changes with a Markov assumption on the dynamics, TOT models absolute timestamp values. This allows TOT to see long-range dependencies in time to predict absolute time values given an unstamped document [16], and to predict topic distributions given a timestamp. It also helps avoid a Markov model's risk of inappropriately dividing a topic in two when there is a brief gap in its appearance. Time is intrinsically continuous. Discretization of time always begs the question of selecting the slice size and the size is invariably too small for some regions and too large for others. TOT avoids discretization by associating with each topic a continuous sharing over time. Many possible parameterized distributions are possible. Our earlier experiments were based on Gaussian. All the results in this paper employ the data sharing for which the time range of the data used for parameter estimation is normalized to a range from 0 to 1. Another possible choice of bounded distributions is the Kumaraswamy distribution [17]. Double-bounded distributions are appropriate because the training data are bounded in time. If it is necessary to predict in a small window into the future the bounded region can be extended, yet still estimated based on the data available up to now. Topics over Time is a generative model of timestamps and the words in the timestamped documents.

A. Algorithm 1 Inference on TOT

```
initialize topic assignment randomly for all tokens
for item = 1 to Niter do
    for d = 1 to D do
        for w = 1 to Nd do
            draw zdw from $P(zdw|w, t, z{-}dw, \alpha, \beta, \Psi)$
            update nzdww and mdzdw
        end for
    end for
    for z = 1 to T do
        update $\psi z$
    end for
end for
compute the posterior estimates of $\theta$ and $\varphi$
```

The above generative model find the data in which there is a timestamp associated with each word. When fitting our model from typical data, each training document's timestamp is copied to all the words in the document. generative model, this process would generate different time stamps for the words within the same document. In this sense, thus, it is formally a deficient generative model, but still remains powerful in modeling large dynamic text collections.

B. Hierarchical Clustering with Constraints

Our algorithm is considered as hierarchical agglomerative clustering with up-to-one mapping constraints. As input takes in three arguments: a list of topic models, a topical similarity measure, and a matching criterion. As output, it generates a list of topical groups, where each group contains a list of topics with at most one topic from each model. At initialization, we create a topical group for every topic in every model. We then iteratively merge the two most similar groups based on the user-supplied topical similarity measure, provided that the groups satisfy the user-specified matching criterion and the mapping constraints. When no new groups can be formed, the algorithm terminates and returns a sorted list of final topical groups [18]. During the alignment process, the following two invariants are guaranteed: Every topic is always assigned to exactly one group; every group contains at most one topic from each model.

C. Event Trigger Extraction

The objective of Automatic Content Extraction (ACE) [19] is to develop automated content extraction techniques to support processing of natural language text from a variety of sources such as newswire, broadcast conversation, and weblogs. ACE has a specific task for event extraction from news sources – which defines event triggers as the words (or phrases) in a sentence that specify the occurrence as well as the type of the event. In a standard event extraction task, triggers are extracted at a sentence level to understand the type of event mentioned in the sentence. Our goal is to understand the "best" event type that describes an entire news article. Therefore, we identify which triggers in a news article collectively describe the central event of the article. Typically, a news article is organized as follows – a title or headline which contains a one line overview of the main event; followed by the lead (or first) paragraph of the article which contains a brief description of the main event; then the rest of the article presents details of the central event along with follow-up actions of the main event. Based on this standard flow of a news article [20], we assume that the triggers appearing in the title and lead paragraph of the article are representative of the main event and consequently, form part of the set of all event class triggers.

NEW DETAILS
APPLE SLAMMED FOR DEFYING JUDGE'S ORDER
GMA
PATIENT SINGS FOR POPE DURING HOSPITAL VISIT
LIVE
CNN
LIVE
BREAKING NEWS
PART OF THE WESTBOUND NJ TURNPIKE IS CLOSED IN JERSEY CITY AFTER ACCIDENT
7

Figure 4: frames of news videos, right column: images from Twitter.

5. EXPERIMENTAL RESULTS

In this section, we present the topics finding the TOT model and compare them with topics from LDA. We also demonstrate the ability of the TOT model to predict the timestamps of documents, Many doubling accuracy in comparison with LDA. We furthermore find topics discovered by TOT to be more distinct from each other than LDA topics (as measured by KL Divergence Finally we take TOT can be used to analyze topic co-occurrence conditioned on a timestamp. A hash tag is considered relevant to a video if the tweets and concepts behind the hash tag is closely related to the video content. We have developed a website for annotation in which the annotators can search tweets and photos from a hashtag, and then can decide if the hash tag is relevant for the given video.

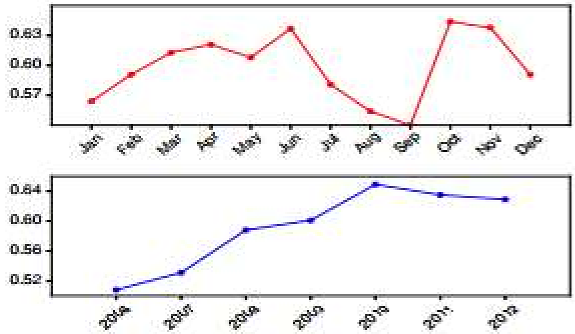


Figure 4: EVENT model for all crops for (upper) and average accuracy across 7 years [2006-2012] (lower)

The event model presented based on the assumption that every news article has a main event, which is usually mentioned in the title and/or lead paragraph of the article.

In the scenario many types of events are discussed in the lead paragraph of the same article, we still preserve the assumption that there is only one main event in the article by choosing the event appearing first among the two events in the lead paragraph, as the main event of the article. This is based on the intuition that the preceding event has more importance and needs to be mentioned before introducing or explaining the other events in the lead paragraph.

6. CONCLUSIONS

Our event-driven predictive model showed a 22% reduction in the RMSE when a standard ARIMA model is incorporated with event information. A separate model was built to predict spikes in the prices the event-based model outperformed the other models, including an LDA based predictive model. Our model demonstrates the value of leveraging the visual and text modalities within both television and social media as they bring disparate but complimentary information for this task. We take automatically leveraging hash tags to place traditional media in its social media context is a useful and important challenge. TOT does not require discretization of time or Markov assumptions on state dynamics. The relative simplicity of our method provides advantages for injecting these ideas into other topic models. One future direction is extending the evaluation to build predictive models for other scenarios such as stock prices, disease outbreaks effects of natural calamities like drought, flood, cyclone etc. The present work has only focused on finding association between events and an observable phenomenon. Another extension of this idea is to identify causing and resulting events of any phenomenon, which can be used for finer analysis and greater clarity in understanding the observed phenomenon.

7. REFERENCES

[1] D. Patel, W. Hsu, and M. L. Lee, “Mining relationships among intervalbased events for classification,” in Proc. SIGMOD, 2008, pp. 393–404.

[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] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed membership models of scientific publications. Proceedings of the National Academy of Sciences, 101(Suppl. 1), 2004.

[5] T. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl. 1):5228–5235, 2004

[6] L. Ballan, M. Bertini, T. Uricchio, and A. Del Bimbo. Data-driven approaches for social image and video tagging. Multimedia Tools and Applications, 74(4):1443–1468, 2015.

[7] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. 2011.

[8] T. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl. 1):5228–5235, 2004.

[9] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In Advances in Neural Information Processing Systems (NIPS) 17, 2004.

[10] Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, (4):32–43, 1999.

[11] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 28–36. ACM, 1998.

[12] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.

[13] M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi, and S. Kumar. Google correlate whitepaper.

[14] A. Hald. On the history of maximum likelihood in relation to inverse probability and least squares. Statist. Sci., 14(2):214–222, 05 1999.

[15] D. Headey and S. Fan. Anatomy of a crisis: the causes and consequences of surging food prices. Agricultural Economics, 39(s1):375–391, 2008.

[16] P. Kumaraswamy. A generalized probability density function for double-

bounded random processes. Journal of Hydrology, 46:79–88, 1980.

[17] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In Proceedings of the 22nd International Conference on Machine Learning, 2005

[18] Richard A. Becker and William S. Cleveland. 1987. Brushing scatterplots. Technometrics, 29(2):127–142. Bernard Berelson. 1952. Content analysis in communication research. Free Pres

[19] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation.

[20] J. E. Engelberg and C. A. Parsons. The causal impact of media in financial markets. J. of Fin, 66(1):67–97, 2011.

ABOUT AUTHORS:

V.LAKSHMI SWATHI is currently pursuing her **MCA** in MCA Department, St.Ann's College Of Engineering& Technology Chirala, A.p. She Received her Bachleor of Science from **ANU**.

G.SUBBA RAO, he received **M.Tech** from **JNTU Kakinada** Presently his working as an Associate Professor in **MCA** St.Ann's College Of Engineering& Technology Chirala, A.p. His research includes data Mining.