Detecting significant events in news corpora

ABSTRACT

Retrospective event identification is the task of identifying and summarizing significant events in time-stamped corpora. In this short paper, we present a simple unsupervised method for identifying significant events in a news corpus based on feature temporal mutual information (I_t) and non-negative matrix factorization (NMF). We evaluate our method using a preliminary test collection and evaluation methodology that can be used for future comparative research. We compare our method to a latent Dirichlet allocation (LDA) approach to better understand the differences between events and topics.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Keywords

Event detection

1. INTRODUCTION

Imagine a Wikipedia editor creating the 'year' page (e.g., 1988) containing a list of important events that occurred during that year. Given a corpus of news articles (or other timestamped collection), it should be possible to show the editor a list of major events that were reported during the time period to facilitate the creation or validation of these pages. Such a system might rank the events by importance, provide estimated dates of event occurrence, and a ranking of candidate Wikipedia pages that describe the associated event.

In this short paper, we present a novel method for retrospective event identification in news corpora based on temporal mutual information [11] and non-negative matrix factorization [9]. We also introduce a preliminary test collection and evaluation methodology based on Wikipedia that can be used for future comparative research. As far as we know, we are the first to propose a method for comparative evaluation in this type of retrospective event identification task.

The results of our method are compared to a baseline implementation[1] of the model described in [5]. Our model outperforms the baseline ...

2. WHAT IS AN EVENT?

The concepts of 'event detection' or 'event extraction' are common in the information retrieval and natural language processing literature. In this paper, we are concerned with the retrospective identification of large-scale events as reported in historical news corpora. A number of studies have used the concepts of "significant," "seminal," or "newsworthy" events to describe such large-scale events. Our notion of event is most closely related to the "seminal events" of the Topic Detection and Tracking (TDT) task, although the task itself is quite different. Examples of large-scale events include major elections, natural disasters, political revolutions, wars, and acts of violence.

3. RELATED WORK

Recent studies in retrospective event identification apply feature-centric models [14, 4, 2, 11, 12] that first identify interesting terms (or features) which are subsequently grouped or combined to represent events.

For example, Fung et al [4] identify temporally interesting features using a series of binomial distributions. The terms are subsequently combined into events using a cost-minimization approach that balances the similarity of the temporal distributions of terms and the co-occurrence of terms in documents.

Weng and Lee [12] use wavelet analysis to identify interesting features and then apply graph partitioning to a graph with terms as nodes and the cross-correlation of term time-series as edge weights. This approach does not account for term co-occurrence in documents.

He, Chang and Lim [5] use spectral analysis to identify interesting features in text corpora. Features are grouped into events using a cost-minimization approach that combines the similarity of the temporal distributions using KL

SIGIR '2016 Pisa, Italy

[1] Although we contacted the authors, we were unable to gain access to the original source code. We acknowledge that our implementation may not be exactly as described in the original article due to some ambiguities in the original model description.

divergence and co-occurrence of terms in documents. We use this model as a baseline for comparison.

Also related to the current study is research in latent semantic analysis (LSA) [3, 6] and topic modeling [1]. Event models can be seen as a special type of temporally-constrained topic models. Our approach is akin to the factor-based LSA model, but we anticipate expanding this work to a generative probabilistic framework in the future.

4. EVENT DETECTION MODEL

Each of the feature-centric models described above includes two basic steps. The first is to identify terms or features that are indicative of an event in the corpus, generally using time-series methods. The second is to combine terms generated by the same underlying event into event models. This is usually done by balancing temporal similarity (e.g., time series correlation) with term co-occurrence in documents.

In this study, we identify temporally interesting terms using the first-order autocorrelation (ACF) of the term time series [7]. We then calculate the temporal mutual information between the top K terms based on ACF and apply non-negative matrix factorization (NMF) to identify latent factors. We interpret the resulting factors as events. The basic process is defined as follows:

First, we construct a temporal index from the document collection that consists of time series for each term in the vocabulary at a particular interval (e.g., hour, day, week):

Definition 1. Feature time series. *The time series of a feature f is defined as the sequence:*

$$y_f = [y_f(1), y_f(2), ..., y_f(T)],$$

where each element $y_f(t)$ is a measure of feature f at time t. For this paper, we use the simple document frequency:

$$y_f(t) = DF_f(t)$$

where $DF_f(t)$ is the number of documents containing feature f at time t.

Next, for each term in the vocabulary, we calculate the first-order autocorrelation (ρ) of the time series.

Definition 2. First order autocorrelation. *Given*

$$\rho = \frac{\sum_{i=1}^{N-1}(y_t - \bar{y}_{(1)})(y_{t+1} - \bar{y}_{(2)})}{\left[\sum_{i=1}^{N-1}(y_t - \bar{y}_{(1)})^2\right]^{1/2}\left[\sum_{i=1}^{N-1}(y_{t+1} - \bar{y}_{(2)})^2\right]^{1/2}}$$

Terms with high temporal dependency will have very high (1) or very low (-1) ρ. For this short paper, we consider only the top K features based on ACF, where K is chosen heuristically.

For all terms with high temporal dependency, we calculate the temporal mutual information $I_t(x;y|t)$ between terms [11].

Definition 3. Temporal mutual information. *Given a timestamp t and pair of terms x and y, the temporal mutual information between x and y in t is defined by:*

$$I_t(x;y|t) = p(x,y|t)\log\frac{p(x,y|t)}{p(x|t)p(y|t)}$$

In this study, $I_t(x;y|t)$ is calculated independently for each interval (as opposed to cumulatively, as described in [11]).

Finally, we construct a symmetric matrix of the maximum I_t between terms and apply NMF [9] to identify latent factors.

Definition 4. Non-negative matrix factorization. *Given a non-negative matrix V, NMF finds matrix factors W and H such that:*

$$V \approx WH$$

We interpret the matrix W as latent event models, using the assigned weights as term weights. The top N models are evaluated as described in a latter section.

5. BASELINE LDA MODEL

As noted above, there is a relationship between event detection and topic modeling, such as latent Dirichlet allocation (LDA) [1]. Some topics are likely generated by events, particularly in news corpora. Even though LDA does not explicitly incorporate time, we expect to find some event-related topics. The problem is how to identify them.

One approach is to consider LDA topics ranked by the first order autocorrelation (ACF) of the topic time-series. The time series of a topic m is defined as the sequence:

$$y_m = [y_m(1), y_m(2), ..., y_m(T)],$$

where each element $y_m(t)$ is a measure of the topic m at time t. For this paper, we use the sum of the document-topic probability from LDA:

$$y_m(t) = \sum_{d \in D} p(m|d,t)$$

We calculate the topic time series ACF as described in the previous section. As in [12], we compare our event detection model to the baseline of LDA-generated topics ranked by topic time series ACF. The top N topic models are evaluated as described in the next section.

6. EVALUATION

Previous studies in feature-centric event detection have not provided reusable test collections or methodologies for comparative evaluation. In general, authors have analyzed the output of proposed models in isolation. Only [12] compares the results of their models to LDA, but only qualitatively. In this section, we outline a preliminary methodology and test collection that can be used for future comparative evaluation.

The primary concern in the evaluation of feature-centric event detection models is how effective the model is at identifying events in the corpus. For this short paper, we propose an evaluation methodology based on Wikipedia and the pooled results of the systems under comparison.

Before evaluation, each system generates a set of N candidate event models based on the test corpus. Each model contains a list of k terms and associated weights.

To develop a ground-truth of *known-events*, we start with events listed in the Wikipedia 'year' pages for the period of the test collection. The known-event list contains an event ID, description, date, and associated Wikipedia URL. With this initial data, the pooled candidate events from the systems are then manually reviewed to identify any events missing from the known-events list. Each candidate event model is compared to the list and, if it is determined by the assessor to represent an event that is missing, is added. At

ID	Date	Description	Wikipedia URL
19	Mar 11	Terrorists execute simultaneous attacks, with bombs in 4 rush-hour trains in Madrid, killing 191 people.	2004_Madrid_train_bombings
42	June 28	The U.S. led coalition occupying Iraq transfers sovereignty to an Iraqi Interim Government.	Iraqi_sovereignty
57	Sept 1	Chechen terrorists take 1,128 people hostage, mostly children, in a school in the Beslan school hostage crisis.	Beslan_school_siege
76	Nov 2	United States presidential election, 2004: Republican incumbent President George W. Bush is declared the winner over his Democratic challenger, U.S. Senator John F. Kerry, in a close election.	United_States_presidential_election,_2004
91	Dec 26	One of the worst natural disasters in recorded history hits Southeast Asia, when the strongest earthquake in 40 years, measuring 9.3 on the Richter scale, hits the entire Indian Ocean region, which generates an enormous tsunami that crashes into the coastal areas of a number of nations.	2004_Indian_Ocean_earthquake_and_tsunami

Table 1: Example entries from the known-events list (Wikipedia 2004)

the end of this process, we have a ground-truth of unique known-events represented as URLs, based on a combination of the Wikipedia year pages and the pooled results from the systems.

For evaluation, we use the candidate event models (i.e., terms and term weights) as queries against an indexed copy of Wikipedia, returning the top k pages. We hypothesize that a high-quality event model will serve as an effective query. For each of the top N candidate events, we compare the number of known-event URLs returned at each depth (i.e., $1 < k < 10$) and based on the proportion of known events identified in the top N candidate events at each depth.

6.1 Test collection

News articles from the year 2004 in the New York Times Annotated Corpus [10] are used for evaluation. Only articles found in section A (available via document metadata) are considered and summary articles are ignored. Stopwords are removed using the standard Indri list. The resulting collection consists of 25,689 documents.

The "known-events" list is constructed using the process described in the previous section based on the "Events" section of the 2004 Wikipedia year page[2] and the top $N = 50$ candidate events from the two systems. The final "known-events" list contains 140 unique events with associated Wikipedia URLs. Of these events, 95 are from the Wikipedia year page and the remaining 55 were added based on the pooled system results. Selected entries from the known-events list are presented in Table 1. The indexed copy of Wikipedia is from 09/01/2015.

7. RESULTS

The TMINMF model was run using the top 1000 features based on ACF and $k = 50$ factors for NMF. The LDA model was run using the Mallet implementation with $T = 100$ topics, an interval of 10 for hyperparameter optimization, and all other default values. Only the top $N = 50$ LDA topics were considered as events for evaluation, based on topic time series ACF.

The top 5 events from both the TMINMF and LDA methods are presented in Table 2. Outwardly, the models look similar and would be difficult to qualitatively evaluated. Figure 1 compares the TMINMF and LDA methods under four separate conditions.

The first graph shows the number of known-events identified at each URL depth ($k \in 1 : 10$). Looking only at the first URL returned by each system for all events, the LDA model identified 5 known events, whereas the TMINMF method identified 18. Looking at all 10 URLs returned for each event, the LDA model identified 9 known events, while the TMINMF method identified 36. Note, multiple known-event URLs returned for the same candidate event are treated as a single known event. Duplicate events are allowed. (Note: I think this is confusing and will revise to limit to only unique events. The fact is, both LDA and TMINMF return multiple models for the same events).

The second graph shows the number of unique known-event URLs returned at each depth. In this case, the known-event URLs are counted independent of the candidate events – as is a user were looking only at a list of unique Wikipedia URLs.

The third graph shows the number of known-events represented by the first URL returned for each candidate event at each event rank ($N \in 1 : 50$). The fourth graph is the same, but for the top 10 URLs returned for each candidate event.

From these graphs, we can see that the TMINMF model is consistently more effective at identifying known-events than the ACF-ranked LDA topics. This is both in terms of event coverage (i.e., more known events were returned) and in event model quality (i.e., the event models serve as better queries to Wikipedia).

Additionally, the TMINMF approach identifies 3 events that have no representation in Wikipedia at all.

8. CONCLUSIONS

This short paper introduces a simple and novel method for retrospective event detection in news corpora. Previous research in the area of retrospective event detection has relied on isolated qualitative evaluation to assess model effectiveness. In this paper, we introduce a preliminary evaluation methodology that can be used for comparative evaluation. Using this evaluation methodology, we demonstrate that our simple model is effective and outperforms an LDA-based approach.

The TMINMF approach has several strengths. First, it is principled, using a well-known method for factor analy-

[2] https://en.wikipedia.org/wiki/2004#Events

sis (NMF). Second, it has a natural way of weighting terms based on the assigned factor weights. Third, terms can be in multiple event models. Fourth, temporal mutual information provides a principled approach to the measurement of temporal relationships between terms than the cost minimization approaches used in earlier studies.

Future work: Evaluation over a larger corpus (the LDC NYT collection contains articles from 1987-2007). Estimation and evaluation of event dates. Methods for ranking events by importance. Generative model, ala LDA.

9. ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grant No. [blind]. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

10. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. *CIKM '09*, page 523, 2009.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[4] G. Fung, J. Yu, P. Yu, and H. Lu. Parameter free bursty events detection in text streams. *VLDB '05*, pages 181–192, 2005.

[5] Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. *SIGIR '07*, page 207, 2007.

[6] T. Hofmann. Probabilistic latent semantic indexing. *SIGIR '99*, pages 50–57, 1999.

[7] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems*, 25(3):14–es, 2007.

[8] J. Kleinberg. Bursty and hierarchical structure in streams. *SIGKDD '02*, page 91, 2002.

[9] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, (1):556–562, 2001.

[10] E. Sandhaus. The New York Times Annotated Corpus LDC2008T19, 2008.

[11] C. Teng and H. Chen. Event Detection and Summarization in Weblogs with Temporal Collocations. In *LREC'08*, 2008.

[12] J. Weng and B. Lee. Event Detection in Twitter. In *Proc. Fifth International AAAI Conference on Weblogs and Social Media*, number 98, pages 401–408, 2011.

[13] Y. Yang, T. Pierce, and J. Carbonell. A Study of Retrospective and On-line Event Detection. *SIGIR '98*, pages 28–36, 1998.

[14] J. Yi. Detecting buzz from time-sequenced document streams. In *IEEE EEE 2005*, pages 347–352, 2005.

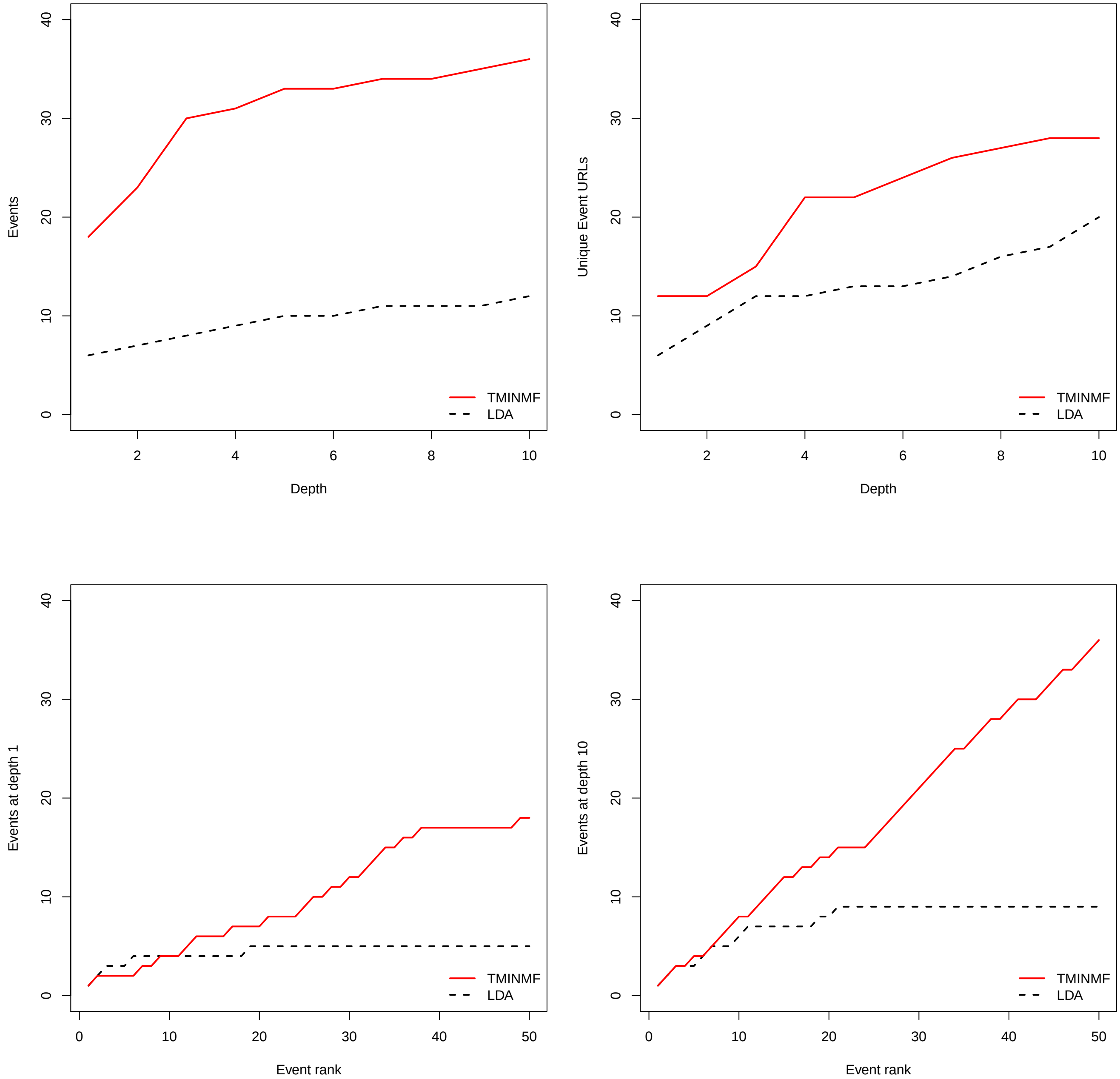

Figure 1: (a) number of known-events at each URL depth, (b) number of unique known-event URLs at each depth (c) number of known events at depth 1 for all ranks, (d) number of known events at depth 10 for all ranks.

ID	**TMINMF**	**LDA**
1	wesley(0.075) clark(0.074) dean(0.055) primary(0.052) howard(0.05) edwards(0.05) kerry(0.049) hampshire(0.048) 2004(0.045) missouri(0.044)	edwards(0.048) dean(0.048) campaign(0.043) dr(0.034) kerry(0.025) candidate(0.024) s(0.018) iowa(0.018) primary(0.016) clark(0.016)
2	taguba(0.036) soldier(0.035) detainee(0.03) prisoner(0.029) karpinski(0.029) ricardo(0.028) fay(0.028) maj(0.028) command(0.027) cellblock(0.027)	prisoner(0.034) military(0.033) abuse(0.026) detainee(0.025) interrogate(0.023) iraq(0.023) prison(0.021) abu(0.018) ghraib(0.018) america(0.013)
3	poll(0.181) voter(0.131) politics(0.131) bush(0.112) election(0.11) campaign(0.102) vote(0.098) candidate(0.066) senator(0.061) kerry(0.056)	kerry(0.135) bush(0.071) campaign(0.046) s(0.038) president(0.028) john(0.018) senator(0.016) cheney(0.013) republican(0.012) democrat(0.011)
4	port(0.044) au(0.04) aristide(0.038) prince(0.033) philippe(0.032) rebel(0.03) haitien(0.028) gona(0.028) politics(0.028) haiti(0.025)	haiti(0.049) aristide(0.037) government(0.022) cuba(0.018) s(0.015) president(0.014) rebel(0.013) chile(0.013) zimbabwe(0.012) country(0.011)
5	sadr(0.05) moktada(0.044) baghdad(0.042) sunni(0.039) militia(0.037) falluja(0.036) occupation(0.03) gen(0.027) struggle(0.026) karbala(0.026)	vote(0.077) election(0.059) voter(0.057) percent(0.03) poll(0.027) ballot(0.022) party(0.021) candidate(0.018) republican(0.016) democratic(0.015)

Table 2: Top 5 events detected by TMINMF and LDA methods