Detecting Public Influence on News Using Topic-aware Dynamic Granger Test

Lei Hou[1], Juanzi Li[1], Xiao-Li Li[2], and Jianbin Jin[3]

[1]Department of Computer Sci. & Tech., Tsinghua University, Beijing 100084, China
[2]Institute for Infocomm Research, A*STAR, Singapore 138632
[3]School of Journalism & Communication, Tsinghua University, Beijing 100084, China
{houl10@mails.,lijuanzi@}tsinghua.edu.cn
xlli@i2r.a-star.edu.sg, jinjb@mail.tsinghua.edu.cn

Abstract. With the rapid proliferation of Web 2.0, *user-generated content (UGC)*, which is formed by the public to reflect their views and voice, presents rich and timely feedback on news events. Existing research either studies the common and private features between news and UGC, or describes the ability of news media to influence the public opinion. However, in the current highly media-user interactive environment, investigating the public influence on news is of great significance to risk and credible management for government and enterprises. In this paper, we propose a novel topic-aware dynamic Granger test framework to quantify and characterize the public influence on news. In particular, we represent words and documents as distributed low-dimensional vectors which facilitates the subsequent topic extraction. Then, a topic-aware dynamic strategy is proposed to transfer news and UGC streams into topic series, and finally we apply Granger causality test to investigate the public influence on news. Extensive experiments on 45 diverse real-world events demonstrate the effectiveness of the proposed method, and the results show promising prospects on predicting whether an event will be properly handled at its early stage.

Keywords: News, User-Generated Content, Influence, Distributed Representation, Granger Causality

1 Introduction

Social media presents rich and timely feedback on news events that take place around the world. According to the report from *Pew Research Center*, 63% of social users from Twitter and Facebook accessed news online, and roughly a quarter of them actively expressed their opinions on daily news through these social applications [2]. The various user-generated content not only fuels the news with different events from different perspectives, but also spurs additional news coverage in the event. On the other hand, reading the social media, understanding and responding to public voice timely and objectively, can help news media promote its influence on the public.

Example. In the event *Asia-Pacific Economic Cooperation* (APEC) 2014 in China, *region cooperation*, *global economic* were the topics supposed to be reported by news media. However, in fact, social media users posted significant amount of comments on *APEC blue* – rare blue sky in Beijing during the summit due to emission reduction campaign directed by Chinese government. Following that, news media quickly followed and paid great attention on this topic which was beyond the original news agenda. We found that 38 of the 176 news articles on Sina were related to the *APEC blue*. Furthermore, how news responds to the public voice has a significant impact on the government credibility. For example, two severe earthquakes struck China in 2014, i.e., *Yunnan* and *Sichuan*. Both reports covered the major topics, but in *Yunnan* earthquake, news media responded the public timely and pictured a comprehensive image of event progress from the perspectives of the public, and thus harvested better support from the public. Therefore, investigating the public influence on news is of great benefit to public opinion management and government credibility improvement.

Related researches on mutually news and UGC stream analysis mainly follows three lines. The first studies event evolution within *individual* news stream [1, 8], e.g., Mei and Zhai adapt PLSA to extract topics in news stream, then identify coherent topics over time, and finally analyze their lifecycle [17]. The second focuses on simultaneously modeling *multiple* news streams, e.g. identifying the characteristics of social media and news media [26] or their common and private features [10, 23]. But both of them pay little attention on the interactions between two streams which could inspire their co-evolution. The last comes from the journalism communication. It applies agenda setting theory [16] to analyze the interactions between different news agenda, and it is often completed via questionnaire survey or manual work on limited events. However, in the era of social media, agenda setting is not a one-way pattern from news to the public, but rather a complex and dynamic interaction.

Detecting the public influence on news poses unique technical challenges: i) most researches use latent topic to model news and UGC, but the traditional word distribution representation [17, 5] suffers from the sparsity problem due to the UGC's short and fragmented characteristics, making it difficult to track topic changes; ii) how to detect the cross-media influence links remains another problem, since the commonly-used measures (e.g., KL-divergence [17, 12]) often leads to heuristic results without statistical explanation.

In this paper, we propose a novel topic-aware dynamic Granger test framework to automatically study the public influence on news media. To address the sparsity problem, we first represent word as low-dimensional word vectors through skip-gram model [19], and further reform word representation via s-parse coding to capture the latent semantic of each dimension. Then we employ Granger causality test [9] to theoretically detect the public influence on news. Particularly, for a pair of topics extracted from UGC and news respectively, we propose a topic-aware dynamic strategy to chronologically split those topic-related documents into disjoint bins with dynamic time intervals, calculate the topic representations based on the documents falling into each bin, and apply the multivariate Granger test to judge if the UGC-to-news influence exists. Finally, we quantify the influence [12] based on the discovered influence links, and validate the influence measures by calculating their correlations with the professional, manual results provided by *China Youth Online.*

The main contributions can be summarized as follows:

- We address problem of analyzing public influence analysis on news through a unified Granger-based framework. Extensive experiments are conducted on 45 real-world events to demonstrate its effectiveness, and results could provide useful guidance on handling public hot topics in event reporting.
- We propose a novel textual feature extraction method. Instead of directly using the popular word2vec, it further maps word and document into a low-dimensional space with each dimension denoting a more compact semantic thus facilitates topic extraction and representation.
- To track the temporal changes of topic pair from news and UGC respectively, we propose a novel topic-aware dynamic binning strategy, splitting both streams into chronological bins to achieve smooth topic representations of each bin.

The rest of the paper is organized as follows. In Section 2, we first define the related concepts and the problem of influence analysis from UGC to news. Section 3 presents our proposed textual feature extraction method and Granger-based influence analysis. Our experimental results are reported in Section 4. Section 5 reviews the related literatures, and finally Section 6 concludes this paper with future research directions.

2 Problem Definition

A particular event often brings forth two correlated streams, namely news articles from newsroom form a news stream and the public voice from different social applications converge into a UGC stream. Both news stream *NS* and UGC stream *US* are text streams, which are defined as follows:

Definition 1. ***Text Stream.*** *A text stream $TS = \langle s_1, s_2, \ldots, s_n \rangle$ is a sequence of documents, where s_i $(i = 1, 2, \ldots, n)$ is associated with a pair $(d_i,\ t_i)$, where d_i is a document comprising a list of words and t_i is the publish time in non-descending order, i.e. $t_i \leq t_{i+1}$.*

It has been shown that news and UGC streams are mutually dependent [24]. *Topic*, which bridges these two different streams, plays an important role. In order to study the cross-stream interactions, we first define topic as follows:

Definition 2. ***Topic.*** *Conceptually, topic z expresses an event related subject/theme within a time period. Mathematically, topic $\mathbf{z}$ is characterized as a vector with each dimension denoting a word feature or a latent aspect. Topic z covers multiple documents (news articles or users comments).*

The interaction between media, public and government has been theoretically studied in journalism communication, e.g., the agenda setting theory[1] evaluated the ability of mass media to influence the salience of topics on the public [15]. Nowadays, the proliferation of social media is changing the way of news diffusion, i.e., the public may inversely affect or even drive the news media. It is useful to explore to what extent the traditional news depends on social media and how long the public influence lasts, thus we condense the following research problem.

Definition 3. ***Analyzing Public Influence on News.*** *Given a news stream NS and a UGC stream US, influence analysis from UGC to news aims to discover a set of influence links* $\{(\mathbf{z}_u, \mathbf{z}_n, \zeta)\}$*, where* $\mathbf{z}_u \in Z_u$ *and* $\mathbf{z}_n \in Z_n$ *are topics extracted from US and NS respectively, and* $\zeta \in \{0, 1\}$ *indicates whether* $\mathbf{z}_u$ *influences (or contributes to)* $\mathbf{z}_n$*.*

From the definition above, topic representation and extraction, influence detection are two major steps to complete the novel task. As mentioned in the introduction, existing methods suffer from various technical deficiencies, i.e., s-parse representation and lack of theoretical foundation. To tackle these issues, we put effort on the following two problems: i) given news and UGC streams, properly represent the documents and extract latent topics from both stream-s; ii) given a topic pair $(\mathbf{z}_u, \mathbf{z}_n)$, determine if there exists a causality link and provide a statistical evaluation on how $\mathbf{z}_u$ contributes to $\mathbf{z}_n$.

3 Our Approach

In this paper, we propose a topic-aware dynamic Granger-based method to automatically detect the influence from UGC to news. Specifically, we develop a text representation method to better represent news and UGC in a low-dimensional space and extract their corresponding topics (Section 3.1). We incorporate temporal information to transform news and UGC topics into serialized representations and apply Granger causality test to detect the public influence on news (Section 3.2).

3.1 Text Representation and Topic Extraction

Text representation and topic extraction aims to properly represent the documents in *NS* and *US* and extract topics Z_n and Z_u. However, traditional TF-IDF representations suffer problems of the curse of dimensionality and feature independence assumption in dealing with the short and fragmented UGC. These methods often ignore the semantic relationships among word features which leads to document sparse representation with many zero features values.

To alleviate the sparse representation, many methods have been proposed to unveil the hidden semantics of words, such as topic models (e.g., LDA [3]) and external knowledge enrichment (e.g., ESA [7]). However, topic models rely

[1] https://en.wikipedia.org/wiki/Agenda-setting_theory

much on the word co-occurrence that cannot be accurately computed with short texts, while ESA requires plenty of high-quality knowledge, which is often not available in practice. In this paper, we propose a novel textual feature extraction pipeline, which gradually maps word and document into a low dimensional space where each dimension represents a unique semantic meaning. It consists of the following three steps:

Word vectorization. Word is the basic element in text, so we first transform words into continuous *low-dimensional* vectors. Let V denote the vocabulary in *NS* and *US*, we employ skip-gram model [19] to learn a mapping function: $V \rightarrow \mathbb{R}^M$, where $\mathbb{R}^M$ is a M-dimensional vector. Specifically, given a document $s \in NS \cup US$ associated with word sequence $\langle w_1, w_2, \ldots, w_W \rangle$, skip-gram model maximizes the co-occurrence probability among words that appear within a contextual window k:

$$\max_{\mathbf{w}} \frac{1}{W} \sum_{i=1}^{W} \sum_{j=i-k, j\neq 0}^{j=i+k} \log p(w_j|w_i) \tag{1}$$

The probability $p(w_j|w_i)$ is formulated as:

$$p(w_j|w_i) = \frac{\exp(\mathbf{w}_j^{\mathrm{T}} \mathbf{w}_i)}{\sum_{l=1}^{V} \exp(\mathbf{w}_l^{\mathrm{T}} \mathbf{w}_i)} \tag{2}$$

where $\mathbf{w}_i \in \mathbb{R}^M$ is the M-dimensional representation of w_i.

Mid-level feature learning. Intuitively, the document representation can be achieved via word vector composition. However, each dimension in word vector represents a latent meaning and word semantic scatters over almost all dimensions, simple composition of individual word vectors ignores the potential correlation between dimensions [20]. To prevent possible information loss by simple composition, we reconstruct each word vector into a mid-level feature [4], where each dimension represents a unique dense semantic. In other words, we learn a $\mathbb{R}^M \rightarrow \mathbb{R}^N$ mapping, and it is typically a sparse coding problem, whose objective is:

$$\min_{\mathbf{W}^*, \mathbf{D}} \sum_{i=1}^{V} \|\mathbf{w}_i - \mathbf{D}\mathbf{w}_i^*\|_2^2 + \lambda \|\mathbf{w}_i^*\|_1 \tag{3}$$

where $\mathbf{w}_i \in \mathbb{R}^M$ is the vector obtained in word vectorization; $\mathbf{w}_i^* \in W^* \subseteq \mathbb{R}^N$ is the N-dimensional sparse representation ($N > M$); $\mathbf{D}$ is an $M \times N$ matrix with each column denoting a dense sematic; $\| \cdot \|_1$ denotes the ℓ_1-norm of input vector; $\lambda > 0$ is a hyperparameter controlling the sparsity of the result representation, i.e., larger (or smaller) λ induces more (or less) sparseness of $\mathbf{w}_i^*$. Because the vocabulary V usually contains tens of thousands of words, optimization of the non-convex problem would be very time consuming.

To efficiently solve the problem, we apply a two-step approximation method. Firstly, we learn the matrix $\mathbf{D}$ offline. We cluster the learned word vectors into N clusters through K-means where each cluster denotes a compact semantic, and use the cluster centers as the columns of D. Secondly, based on the assumption that locality is more essential than sparsity [22], we select the K-nearest neighbors in D for each word $\mathbf{w}_i$ based on Euclidean distance, and then adopt the Locality-constraint Linear Coding (LLC) to learn its transformation $\mathbf{w}_i^*$:

$$\begin{aligned}&\min_{\mathbf{W}^*}\sum_{i=1}^{V}\|\mathbf{w}_i-\mathbf{B}_i\mathbf{w}_i^*\|_2^2+\lambda\|\mathbf{w}_i^*\|_2^2\\&s.t.\mathbf{1}^{\mathrm{T}}\mathbf{w}_i^*=1,\forall i\end{aligned}\tag{4}$$

where $\mathbf{B}_i$ is the K-nearest neighbors to $\mathbf{w}_i$ in $\mathbf{D}$. The problem could be solved analytically by:

$$\begin{aligned}\widehat{\mathbf{w}_i^*}&=(\mathbf{V}_i+\lambda\mathbf{I})\backslash\mathbf{1}\\\mathbf{w}_i^*&=\widehat{\mathbf{w}_i^*}/\mathbf{1}^{\mathrm{T}}\widehat{\mathbf{w}_i^*}\end{aligned}\tag{5}$$

where $\mathbf{V}_i=(\mathbf{B}_i-\mathbf{1}\mathbf{w}_i^{\mathrm{T}})^{\mathrm{T}}(\mathbf{B}_i-\mathbf{1}\mathbf{w}_i^{\mathrm{T}})$ denotes the covariance matrix.

Document representation and topic extraction. We employ spatial pooling to represent each document as a N-dimensional vector $\mathbb{R}^N$ based on the learned sparse word vectors. Given a document s_i consisting W words with vector representations $\mathbf{w}_i^*, i=1,2,\ldots,W$, we try two different pooling functions to obtain the document representation $\mathbf{s}_i$:

$$\underbrace{s_{ij}=\frac{1}{W}\sum\nolimits_{k=1}^{W}|\mathbf{w}_{kj}^*|}_{average}\quad or\quad \underbrace{s_{ij}=\sqrt{\frac{1}{W}\sum\nolimits_{k=1}^{W}{\mathbf{w}_{kj}^*}^2}}_{square\ root}\tag{6}$$

where $\mathbf{s}_i$ denotes the final representation of s_i and $s_{ij}|_{j=1}^{N}$ is the j-th entry. Note that different pooling functions assume the underlying distributions differently. Once completing the document representation, we feed the news and comment vectors into K-means algorithm separately to obtain topic sets Z_n and Z_u. The achieved topics have more compact distributed representations than TF-IDF, which is convenient to further computation and analysis.

3.2 Topic Influence Detection

Topic influence detection analyzes the relationship between news and UGC topics, which behaves as inter-stream *influence*. Normally, KL-divergence is employed to evaluate topic transition within news stream [17, 13] or topic interaction across streams [12], but the idea is heuristic and results are often restricted within a too short time period to track the topic evolution.

Therefore, we perform the influence detection in a more theoretical way through Granger causality test[2]. Its basic idea is that a *cause* should be helpful in predicting the future values of a time series, beyond what can be predicted solely based on its own historical values [9]. That is to say, a time series x is

[2] http://en.wikipedia.org/wiki/Granger_causality

to Granger cause another time series y, if and only if regressing for y in terms of both past values of y and x is statistically significantly more accurate than regressing for y in terms of past values of y only.

Granger-based influence detection. In this paper, Granger causality analysis is performed on two topics $z_u \in Z_u$ and $z_n \in Z_n$ to test whether z_u is the Granger cause of z_n.

In the previous subsection, we achieve the news and UGC topic sets and their associated documents, but the Granger causality test requires two time series. So we need to turn topics in Z_n and Z_u into time-varying topic series. For each $\mathbf{z} \in Z_n \cup Z_u$, we need to represent it as $\langle \mathbf{z}^t \rangle_{t=1}^T$ where $\mathbf{z}^t$ is the status of topic z at the t-th interval and T is the size of time intervals. A straightforward way is to partition both streams into disjoint slices with *fixed* time intervals (e.g., one day), i.e., equal-size binning. An alternative is equal-depth binning, i.e., evenly partitioning all documents into T bins. For an obtained partition $\langle S^t \rangle_{t=1}^T$, the representation of topic z at the t-th bin $\mathbf{z}^t$ could be simply computed via averaging the related document vectors within that bin:

$$\mathbf{z}^t = \frac{1}{|S_z^t|} \sum\nolimits_{s_z \in S^t} \mathbf{s} \tag{7}$$

where S_z^t denotes the documents within t-th bin that are related to topic z.

Once we get the time-varying representations of two target topics $\langle \mathbf{z}_n^t \rangle_{t=1}^T$ and $\langle \mathbf{z}_u^t \rangle_{t=1}^T$, we first fit two vector autoregressive models (VAR) over these two series:

$$\mathbf{z}_n^t = a_0 + \sum_{i=1}^{q} a_i \mathbf{z}_n^{t-i} + \mathbf{r} \tag{8}$$

$$\mathbf{z}_n^t = a_0 + \sum_{i=1}^{q} (a_i \mathbf{z}_n^{t-i} + b_i \mathbf{z}_u^{t-i}) + \mathbf{r}_u \tag{9}$$

where (8) predicts a news topic $\mathbf{z}_n^t$ at time stamp t purely based on its historical values, i.e., $\mathbf{z}_n^{t-i}$, while (9) considers the historical values from both news and UGC streams, i.e., $\mathbf{z}_n^{t-i}$ and $\mathbf{z}_u^{t-i}$, for prediction; q is a predefined maximum lag to measure how long the influence lasts; $\mathbf{r}_u$ and $\mathbf{r}$ denote the residuals with/without considering the topic z_u.

Then, to test whether or not (9) results in a better regression than (8) with statistical significance, we apply an F-test (some other similar tests could also be chosen). More specifically, we calculate the residual sum of squares RSS and RSS_u, based on which we obtain the F-statistic:

$$F = \frac{(RSS - RSS_u)/q}{RSS_u/(n - 2q - 1)} \sim F(q, n - 2q - 1) \tag{10}$$

Given a confidence coefficient α, we say $\mathbf{z}_u$ Granger causes $\mathbf{z}_n$ if F is greater than a predefined F_α, i.e. $\zeta = 1$ as defined in Section 2, and otherwise $\zeta = 0$.

However, both streams, especially news articles, are often generated nonuniformly. The equal-size binning performs poorly on such streams since it produces many empty intervals without any news or comments, and the equal-depth binning often leads to extremely unbalanced time spans. Either empty interval or unbalanced spans has side effect on Granger test, making it failed or meaningless. **Topic-aware Dynamic Granger Test.** To address the uneven distribution problem, we propose a topic-aware dynamic binning strategy to partition both streams into several disjoint intervals. The motivation for *topic-aware* is that: different topics follow their unique patterns and show various distributions along timeline and the Granger causality test actually processes a topic pair rather than the whole streams at one time, thus one partition only need to deal with documents within target topics from news and UGC respectively. And the *dynamic binning* aims to alleviate problem of the uneven distribution. Let S_z denote the streaming documents associated with topic z, $\langle S_z^t \rangle_{t=1}^T$ is a partition result, we define the following two types of *dispersion*:

- dis_{amount}: the difference between the largest and the smallest bin size with bin size is defined as the number of contained documents;
- dis_{span}: the difference between the largest and the smallest time span.

Our objective is to balance these two dispersions, namely,

$$\min_{\langle S_z^t \rangle_{t=1}^T} dis_{amount} + dis_{span} \quad s.t. |S_z^t| > 0, \forall t \tag{11}$$

Due to the extremely unbalanced volume of news and comments, we perform the optimization on news stream and the comments just follow. The problem could be solved efficiently using dynamic programming (where *dynamic* comes) and the best solution is always available [13].

4 Experiments

In this section, we first briefly introduce our datasets, and then present the detailed experimental results on topic extraction, topic influence detection and further analysis.

4.1 Dataset Description

To evaluate the effectiveness of the proposed methods, we prepare the following two kinds of datasets:

Datasets from Hou's paper [12]. They are composed of five datasets containing four international events: *the Federal Government Shutdown* in both Chinese and English (cFGS/eFGS), *Jang Sung-taek's* (Jang), *The Boston Marathon Bombing* (Boston) and *India Election* (India). They are collected from influential news portals and social media platforms (i.e., Sina, New York Times, Twitter), and the detailed statistics are summarized in Table 1. These datasets are used to evaluate the effectiveness of our topic extraction and influence detection.

Table 1. Datasets: duration, numbers of comments and news articles, max and average number of comments per news

Dataset	Days	#Comments	#News	Comments/News	
				max	avg
cFGS	35	12,995	97	7,818	134
Jang	43	3,291	84	467	39.2
eFGS	53	17,295	136	1,112	127
Boston	46	7,521	211	518	29.4
India	66	4,723	88	113	53.7

Datasets from China Youth Online (CYOL[3]). CYOL is one of the biggest and leading public opinion analysis website in China. It monthly publishes opinion index based on questionnaire surveys from experts and scholars, civil servants, media people, opinion leaders and ordinary Internet users. The index includes five well-defined metrics: *information coverage*, *activeness*, *response arrival rate*, *response recognition rate* and *satisfactory.* For each event reported by CYOL in 2014, we crawled the news articles and comments from Sina[4] if there existed a corresponding special issue. Finally, we collected 40 events, and for each event, there are 140 news articles and 12,849 comments on average. Due to the space limit, the detailed statistics and data will be published later. We incorporated these datasets and published opinion index to evaluate the influence measures that are automatically calculated based on our approach.

4.2 Results for Topic Extraction

In this section, we report the evaluation on text representation and topic extraction, including the experiment setup (settings, baselines and metrics), comparison results and the parameter analysis.

Settings. We use the gensim[5] implementation of word2vec to learn word vectors with $M = 200$, and K-means to generate the transform matrix D with $N = \{128, 256, 512, 1024\}$. For mid-level feature learning, we apply LLC with various K-nearest neighbors, with $K \in \{1, 5, 10, 50\}$. The parameter λ is set to be $1e^{-4}$ as the author suggested.

Baselines. We use DeepDoc to denote our proposed text representation method, and compare it with TF-IDF based method (TF-IDF) and state-of-the-art topic models on news and UGC, i.e., Document Comment Topic Model (DCT) [11] and Cross Dependence Temporal Topic Model (CDTTM) [12].

Metrics. As for the evaluation metrics, we calculate the inner/inter-cluster distance for all topics. The inner-cluster distance (*inner*) is defined as the average distance between documents within topic, and a smaller value indicates a compact cluster. The inter-cluster distance (*inter*) is the average distance from one

[3] http://yuqing.cyol.com/

[4] http://search.sina.com.cn/?t=zt

[5] http://radimrehurek.com/gensim/models/word2vec.html

topic to all the other topics, and a larger value indicates a better result. We also calculate their relative ratios (*ratio*) where a bigger value shows better performance.

Comparison Results. Table 2 presents the comparison results, from which we can see: i) macroscopically, our proposed DeepDoc outperforms three baselines consistently, and DCT is more steady than other methods while the TF-IDF representation obtains the worst performance. ii) under this measurement, CDTTM is not so sensitive to the stream distribution as described in [12], and DeepDoc does not have this problem as we do not include temporal information in clustering.

Table 2. Results for topic extraction: *inner* and *inter* stand for the average inner/inter-cluster distances, and they are related to the dimension of document representation; *ratio* is calculated through dividing the *inter* by *inner* to measure the clustering performance, and a larger *ratio* indicates a better clustering result

Dataset	**TF-IDF**			**DCT**			**CDTTM**			**DeepDoc**		
	inner	*inter*	*ratio*	*inner*	*inter*	*ratio*	*inner*	*inter*	*ratio*	*inner*	*inter*	*ratio*
cFGS	35.86	92.30	2.574	.5211	1.517	2.911	.5406	1.631	3.017	4.091	15.70	3.838
Jang	17.94	40.28	2.245	.5412	1.596	2.949	.5333	1.697	3.182	4.024	15.19	3.774
eFGS	41.13	120.8	2.938	.5285	1.503	2.844	.5083	1.532	3.014	4.973	19.48	3.917
Boston	25.37	53.48	2.108	.5271	1.529	2.901	.5084	1.509	2.968	3.966	16.60	4.185
India	19.91	37.69	1.893	.5986	1.433	2.394	.6095	1.439	2.361	2.808	8.806	3.136

Parameter Analysis. Then the Boston dataset is chosen to investigate the effects of the number of neighbours, pooling function and the number of matrix columns, and the results are presented in Table 3. We have the following conclusions:

- **Number of neighbours (K).** Regardless of various other settings, generally small number of neighbors leads to better clustering results. This is a promising finding, because the smaller the number of neighbors used (i.e., the more sparse the codes are), the faster the computation will be run, and the less the memory will be consumed.
- **Pooling function.** Different choices of pooling functions lead to very different clustering results. The root mean square pooling achieves better performance under almost every settings than average pooling, and the smaller the code sparseness (larger K and smaller matrix), the gap between these two pooling functions is more significant.
- **Number of matrix columns.** It actually denotes the dimensions of transformed space. Intuitively, if the number of dimension is too small, the mid-level representation will lose discriminative power, but words from the same category of documents will be less similar if the size is too big. Here, we mainly focus on the trade-off between reasonably smaller and bigger size. As can be seen from the results, larger size leads to better results when

$K > 10$, while it is likely that smaller matrix is sufficient under higher level of sparseness.

Table 3. Clustering results on *Boston* dataset with various number of neighbors (K), pooling functions and number of matrix columns (N).

#Neighbors	**Pooling Func.**	**N**			
		128	256	512	1,024
K = 1	Sqrt	4.1608	4.1625	3.9000	3.8275
	Avg	2.9550	2.9325	2.2900	1.9258
K = 5	Sqrt	4.0108	4.0925	4.1850	4.0825
	Avg	3.2475	3.2400	2.8442	2.5242
K = 10	Sqrt	4.0108	3.7325	3.7583	4.0033
	Avg	3.2475	3.5092	3.4850	3.5658
K = 50	Sqrt	2.6892	3.0642	3.2858	3.6292
	Avg	2.6267	2.8575	3.1892	3.3625

4.3 Results for Topic Influence Detection

To evaluate our proposed topic-aware dynamic Granger test method (TDG), we perform three series of experiments, namely, 1) the overall comparison with KL-divergence based method in [12], 2) the comparison of different binning methods, and 3) the effect of the maximum lag.

TDG – KL divergence. Hou *et al.* evaluated their method on manually labeled data, and it achieved comparable results to the human annotation. To make the comparison fair, we compare the Granger results with $\alpha = 0.9/0.8$ with their top 10%/20% links (Hou *et al.* included links with distance less than the median value). Through manual evaluation, the Granger test achieves 94% precision while KL gets only 82%, indicating our method significantly outperform theirs. This comes as no surprise because their KL-divergence based method only finds similar patterns in the other stream (it assumes *similar* topics share similar patterns along timeline which may not hold) while our proposed Granger based method discovers the most useful topics in UGC that contribute to predicting the target news topic and thus are more likely to influence the news.

Dynamic binning – Equal size binning. Table 4 shows the number of detected Granger causal links when different time split methods are applied. We can find that, i)*equal-size*: the equal-size binning gets the worst performance because the streams (especially news stream) distribute nonuniformly and it often leads to zero vectors for bins with on documents. Though mean linear interpolation is employed to deal with the zeros, the results are still not so satisfactory. ii)*dynamic*: dynamic binning optimize (11) over whole news stream without distinguish topics. It can handle the uneven distributed streams to some extent, thus finds more influence links. iii)*topic*: since our proposed method tests a pair of topics every time and different topics may follow different patterns, while the dynamic binning is applied on the whole streams, thus it might not perform well on different topic pairs. Therefore, the topic-aware binning further improves the performance.

Table 4. Granger causality links with different time split methods (0.8 and 0.9 are confidence coefficients)

Dataset	*equal-size*		*dynamic*		*topic*	
	0.8	0.9	0.8	0.9	0.8	0.9
cFGS	1	0	4	2	6	4
Jang	1	0	5	2	6	3
eFGS	2	0	5	2	7	4
Boston	1	0	4	2	8	5
India	1	0	3	2	7	4

How long the influence lasts. To choose a proper maximum lag q (i.e., how many historical values are included in the regression), we select five topic pairs to conduct Granger causality test with maximum lag ranging from 1 to 10, and determine the proper value that achieves the best F statistics (divided by $F_{0.8}$ due to the different time spans). Table 5 shows the results from 3 to 7, we observe that the $F/F_{0.8}$ increases initially until $q = 5$ to reach stable status. We therefore execute all the Granger test with q set as 5. Note that here $q = 5$ does not mean 5 days since *topic aware binning* is used for stream split, and actually the average time difference is about 3.2 days, which tells us that news and UGC in the previous 3 days have much more influence on the current news report.

Table 5. F-statistic with maximum lag (q): $F/F_{0.8}$ denotes the average values of the 5 selected topic pairs.

q	3	4	5	6	7
$F/F_{0.8}$	1.704	2.115	2.493	2.487	2.435

4.4 Influence Usage Analysis

This experiment exploits whether our automatically obtained results are consistent with the objective CYOL public opinion index. Specifically, with our achieved influence links $\{(z_u, z_n)\}$, we quantify the public influence on news through news response *rate(NRR)*, *promptness(NRP)* and *effect(NRE)* as defined in [12]. Their comparable measures in *CYOL Public Opinion Index* are *information coverage(IC)*, *response activity(RA)* and *satisfactory(SA)*. We compute three correlation coefficients for the 3 pairs of measures NRR-IC, NRP-RA, and NRE-SA respectively, and higher correlations indicates better results. For comparison, we use LDA + KL-divergence, Hou's methods(CDTTM + KL) as our baselines. We further try to only use first half of the event data for analysis (Ours$^{\frac{1}{2}}$) to test whether it is helpful in predicting the future influence. Table 6 shows the comparison results.

Table 6. Influence usage results.

Methods	**Correlation Coefficient**		
	NRR-IC	*NRP-RA*	*NRE-SA*
LDA+KL	0.6573	-0.5832	0.6029
Hou(CDTTM+KL)	0.6814	-0.5933	0.6157
Ours(DeepDoc+TDG)	0.7232	-0.6419	0.6483
Ours$^{\frac{1}{2}}$	0.7092	-0.6254	0.6085

As shown in Table 6, our method achieves higher correlations with the CYOL measures than other two methods. Furthermore, we notice that only using the first half of event data, our method can achieve comparable results with those on all data. This implies that it can be used on predicting whether an event could be handled properly at early stage.

Case review. Now we review the events mentioned in the introduction. The APEC 2014 summit shows a good example that the social media can influence news media. Besides *APEC blue*, we identify another topic beyond the scheduled ones, i.e., *tourist.* It actually covered the part-time activities of the dignitaries Mrs, especially their clothing. The news media started to report the part-time activities causally. However, the public was very enthusiastic about the Mrs' tourist and discussed a lot about their clothing. To satisfy people's curiosity, news presented systematic introduction of the first lady's activities and dress. Then, we compare the news response in the two earthquakes: both reports covered the major topics — both NRRs are pretty high (*Yunnan* 84% and *Sichuan* 82%); but in *Yunnan* earthquake, news media responded the public more timely — the NRP in *Yunnan* is much smaller than that in *Sichuan*, roughly 0.8 day v.s. 1.4 days. The final satisfactory shows that it is very important to properly handle the heatedly-discussed topics. Our analysis could summarize about which topic that news should response at what time, thus benefits the public opinion management.

5 Related Work

Our work is related to three lines of research as follows:

5.1 Distributed Text Representation

Representing words in continuous vector space has been an appealing pursuit since 1986 [25]. Recently, Mikolov *et al.* developed efficient method to learn high quality word vectors [19], and a host of follow-up achievements have been made on phrase or document representation, such as paragraph-to-vector [20]. Different from these attempts, we are inspired to borrow the state-of-the-art feature extraction pipeline in computer vision [4] to represent word and document in a new space where each dimension denotes a more compact semantic than directly using word2vec.

5.2 Social News Analysis and Topic Evolution

The proliferation of social media encourages researchers to study its relationship between traditional news media, e.g., Zhao *et al.* employed Twitter-LDA to analyze Twitter and New York Times and found Twitter actively helped spread news of important world events although it showed low interests in them [26]. Petrovic *et al.* examined the relation between Twitter and Newsfeeds and concluded that neither streams consistently lead the other to major events [21]. Besides the common and specific characteristics of news and social media, we pay more attention on the cross-stream interaction.

As for the topic evolution, Mei *et al.* solved the problem of discovering evolutionary theme patterns from single text stream [17], Hu *et al.* modeled the topic variations and identified the topic breakpoints in news stream [13]. Wang *et al.* aimed at finding the burst topics from coordinated text streams based on their proposed mixture model [24]. Lin *et al.* formalized the evolution of a topic and its latent diffusion paths in social community as a joint inference problem, and solved it through a mixture model (for text generation) and a Gaussian Markov Random Field (for user-level social influence) [14]. In this paper, we study the interplay of news and UGC within specific events, trying to analyze the cross-media influence and figure out how they co-evolve over time.

5.3 Agenda Setting and Granger Cauality

Agenda-setting is the creation of public awareness and concern of salient issues by the news media. Mccombs and Shaw discussed the function of mass media in agenda setting [16] in 1972. Many researchers studied the interactions between public agenda and news agenda, e.g., Meraz employed time series analysis to measure the influence in political blog and news media [18]. Our work falls into the second-level agenda-setting (also called attribute agenda-setting), and the major advantage of our framework is that, the attributes are predefined and we extract the latent topics automatically.

The Granger causality test [9] is a statistical hypothesis test for determining whether one time series is useful in forecasting the other one. It has been utilized in many areas for causality analysis or prediction, e.g., [6] adapted it to model the temporal dependence from large-scale time series data [6]; Chang *et al.* used it in Twitter user influence modeling. In this paper, we apply the agenda-setting theory and multivariate Granger test to automatically analyze how the social media influence traditional news.

6 Conclusion

In this paper, we analyze the public influence on news through a Granger-based framework: first represent words and documents in distributed low-dimensional space and extract topics from news and UGC streams, then dynamically split streams to achieve changing topic representations on which we employ Granger causality test to detect influence links. Experiments on real-world events demonstrate the effectiveness of the proposed methods and the results show good prospects on predicting whether an event could be properly handled.

It should be note that Granger test attempts to capture an interesting aspect of causality, but certainly is not meant to capture all, e.g., it has little to say about situations in which there is a hidden common cause for the two streams. In the future work, we will try to address the important but challenging issue.

Acknowledgement

The work is supported by 973 Program (No. 2014CB340504), NSFC-ANR (No. 61261130588), and NSFC key project (No. 61533018), Tsinghua University Initiative Scientific Research Program (No. 20131089256) and THU-NUS NExT Co-Lab. Thank Prof. Chua Tat-Seng, Dr. Hanwang Zhang and Xindi Shang from National University of Singapore for discussion on text representation.

References

1. Ahmed, A., Xing, E.: Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. pp. 20–29 (2010)
2. Barthel, M., Shearer, E., Gottfried, J., Mitchell, A.: The evolving role of news on twitter and facebook. Tech. rep., Pew Research Center (July 2015)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research pp. 993–1022 (2003)
4. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2559–2566 (2010)
5. Chang, Y., Wang, X., Mei, Q.Z., Liu, Y.: Towards twitter context summarization with user influence models. In: Proceedings of the 6th ACM international conference on Web search and data mining. pp. 527–536 (2013)
6. Cheng, D., Bahadori, M.T., Liu, Y.: Fblg: A simple and effective approach for temporal dependence discovery from time series data. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 382–391 (2014)
7. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In: AAAI. vol. 6, pp. 1301–1306 (2006)
8. Gohr, A., Hinneburg, A., Schult, R., Spiliopoulou, M.: Topic evolution in a stream of documents. In: the 9th SIAM International Conference on Data Mining. pp. 859–872 (2009)

9. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society pp. 424–438 (1969)
10. Hong, L., Dom, B., Gurumurthy, S., Tsioutsiouliklis, K.: A time-dependent topic model for multiple text streams. In: Proceedings of the 17th ACM International Conference on Knowledge Discovery in Data Mining. pp. 832–840 (2011)
11. Hou, L., Li, J., Li, X.L., Qu, J., Guo, X., Hui, O., Tang, J.: What users care about: a framework for social content alignment. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. pp. 1401–1407 (2013)
12. Hou, L., Li, J., Li, X.L., Su, Y.: Measuring the influence from user-generated content to news via cross-dependence topic modeling. In: Proceedings of the 20th International Conference on Database Systems for Advanced Applications. pp. 125–141 (2015)
13. Hu, P., Huang, M., Xu, P., Li, W., Usadi, A.K., Zhu, X.: Generating breakpoint-based timeline overview for news topic retrospection. In: Proceedings of the 11th IEEE International Conference on Data Mining. pp. 260–269 (2011)
14. Lin, C.X., Mei, Q., Han, J., Jiang, Y., Danilevsky, M.: The joint inference of topic diffusion and evolution in social communities. In: Proceedings of the 11th IEEE International Conference on Data Mining. pp. 378–387 (2011)
15. Lippmann, W.: Public opinion. Transaction Publishers (1946)
16. McCombs, M., Shaw, D.: The agenda-setting function of mass media. Public opinion quarterly pp. 176–187 (1972)
17. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In: Proceedings of the 11th ACM International Conference on Knowledge Discovery in Data Mining. pp. 198–207 (2005)
18. Meraz, S.: Is there an elite hold? traditional media to social media agenda setting influence in blog networks. Journal of Computer-Mediated Communication 14(3), 682–707 (2009)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the ICLR (2013)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
21. Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., Shrimpton., L.: Can twitter replace newswire for breaking news? In: Proceedings of the 7th international AAAI Conference on Weblogs and Social Media (2013)
22. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3360–3367 (2010)
23. Wang, X., Zhang, K., Jin, X., Shen, D.: Mining common topics from multiple asynchronous text streams. In: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. pp. 192–201 (2009)
24. Wang, X., Zhai, C., Hu, X., Sproat, R.: Mining correlated bursty topic patterns from coordinated text streams. In: Proceedings of the 13th ACM International Conference on Knowledge Discovery in Data Mining. pp. 784–793 (2007)
25. Williams, D.R.G.H.R., Hinton, G.: Learning representations by back-propagating errors. Nature pp. 323–533 (1986)
26. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on Information Retrieval. pp. 338–349 (2011)