Measuring the Influence from User-Generated Content to News via Cross-Dependence Topic Modeling

Lei Hou[1](✉), Juanzi Li[1], Xiao-Li Li[2], and Yu Su[3]

[1] Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
houl10@mails.tsinghua.edu.cn, lijuanzi@tsinghua.edu.cn

[2] Institute for Infocomm Research, A*STAR, Singapore 138632, Singapore
xlli@i2r.a-star.edu.sg

[3] Communication Technology Bureau, Xinhua News Agency, Beijing 100803, China
suyu@xinhua.org

Abstract. Online news has become increasingly prevalent as it helps the public access timely information conveniently. Meanwhile, the rapid proliferation of Web 2.0 applications has enabled the public to freely express opinions and comments over news (user-generated content, or UGC for short), making the current Web a highly interactive platform. Generally, a particular event often brings forth two correlated streams from news agencies and the public, and previous work mainly focuses on the topic evolution in single or multiple streams. Studying the inter-stream influence poses a new research challenge. In this paper, we study the mutual influence between news and UGC streams (especially the UGC-to-news direction) through a novel three-phase framework. In particular, we first propose a cross-dependence temporal topic model (CDTTM) for topic extraction, then employ a hybrid method to discover short and long term influence links across streams, and finally introduce four measures to quantify how the unique topics from one stream affect or influence the generation of the other stream (e.g. UGC to news). Extensive experiments are conducted on five actual news datasets from Sina, New York Times and Twitter, and the results demonstrate the effectiveness of the proposed methods. Furthermore, we observe that not only news triggers the generation of UGC, but also UGC conversely drives the news reports.

Keywords: News stream · User-generated content · Cross dependence · Influence · Response

1 Introduction

Nowadays, social media are ubiquitous, offering many opportunities for people to access and share information, to create and distribute content, and to

M. Renz et al. (Eds.): DASFAA 2015, Part I, LNCS 9049, pp. 125–141, 2015.
DOI: 10.1007/978-3-319-18120-2_8

interact with more traditional media [13]. According to the report [14] from *Pew Research Center*, over half of the social users in U.S. access news online, as well as actively express their opinions and comments on daily news, either directly from the online news (comments following the news) or through other services such as blogging, Twitter, which produces the rich user-generated content (UGC). Digital storytelling and consistently available live streaming is fuelling the news with different events from different perspectives [19], indicating the public voice, e.g. their opinions, concerns, requests, debates, reflections, can spur additional news coverage in the event. This comes as no surprise as the main function of the news is to provide updates on the public voice, and the latest measures taken by the involved organizations. Therefore, investigating and responding the public voice is of great benefit to valuable news clues acquisition for news agency, crisis monitoring and management for functional departments.

When a particular event happens, news articles typically form a news stream that records and traces the event's beginning, progression, and impact along a time axis. Meanwhile, the UGC stream is also naturally formed by the public to reflect their views over news reports. These two different streams are highly interactive and inter-dependent. On one hand, the news stream has big influence on the UGC stream as the public posts their comments based on the corresponding news articles and they are typically interested in certain aspects or topics in the news stream. On the other hand, the UGC stream, containing the public opinions, voice and reflection, could potentially influence and even drive the news reports, which is the focus of this paper.

Example. Fig. 1 presents the news and comment streams about *U.S. Federal Government Shutdown* from New York Times (NYT). At the beginning, the news reported the budget bill proposed by the *House of Representatives*, leading to the public debate as well as *Cruz*'s speech. Then the public turned their attention to the following *vote* raised by *the Senate*. They also encouraged the president after the *shutdown*, which might affect the final decision. Meanwhile, the news agencies preferred to report what the public cared most (e.g. *vote*, *insurance*). As such, it is interesting to systematically study how the topics from two correlated streams interact with each other and co-evolve over time.

Recently, many research efforts have been put on topic evolution within news stream, e.g. [1,12]. Morinaga et al. proposed a framework for tracking topical dynamics using a finite mixture model [24]. A representative work from Mei et al. employed adapted PLSA for topic extraction in text streams, KL-divergence for discovering coherent topics over time, and HMM for analyzing the lifecycle in [22]. However, they only detect how the topics evolve but we further explore what factors could drive their evolution. Another line of research focuses on simultaneously modeling multiple news streams, such as mining common and private features [15,29], and identifying the characteristics of the social media and news media via a modified topic model [32]. However, they still did not investigate if the inter-stream influence could lead to their co-evolution.

We apparently expect to study the interactions between news and UGC streams, and address the problem of influence quantification. To the best of our knowledge,

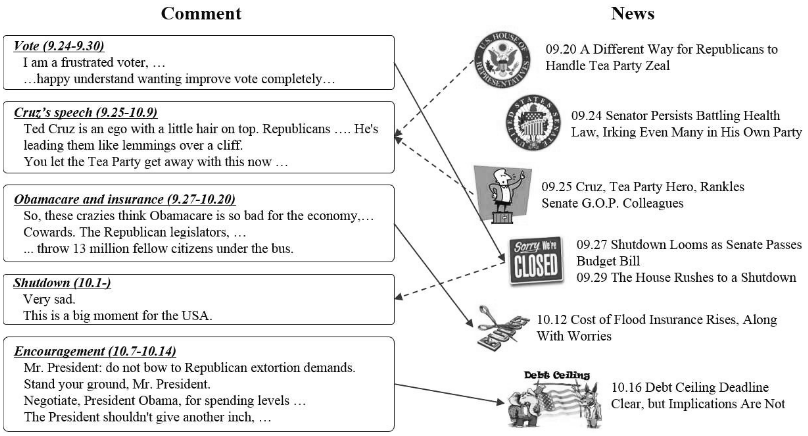
Comment
News
Vote (9.24-9.30)
I am a frustrated voter, …
…happy understand wanting improve vote completely…
Cruz's speech (9.25-10.9)
Ted Cruz is an ego with a little hair on top. Republicans …. He's leading them like lemmings over a cliff.
You let the Tea Party get away with this now …
Obamacare and insurance (9.27-10.20)
So, these crazies think Obamacare is so bad for the economy,…
Cowards. The Republican legislators, …
... throw 13 million fellow citizens under the bus.
Shutdown (10.1-)
Very sad.
This is a big moment for the USA.
Encouragement (10.7-10.14)
Mr. President: do not bow to Republican extortion demands.
Stand your ground, Mr. President.
Negotiate, President Obama, for spending levels …
The President shouldn't give another inch, …
09.20 A Different Way for Republicans to Handle Tea Party Zeal
09.24 Senator Persists Battling Health Law, Irking Even Many in His Own Party
09.25 Cruz, Tea Party Hero, Rankles Senate G.O.P. Colleagues
Sorry We're CLOSED
09.27 Shutdown Looms as Senate Passes Budget Bill
09.29 The House Rushes to a Shutdown
10.12 Cost of Flood Insurance Rises, Along With Worries
Debt Ceiling
10.16 Debt Ceiling Deadline Clear, but Implications Are Not

Fig. 1. News and UGC Interaction in *U.S. Federal Government Shutdown*

it is the first research that focuses on investigating the mutual influence between these two streams. However, the novel task brings new challenges to conventional mining approaches. Firstly, it proves utilizing both streams can significantly benefit the topic discovery process than each individual stream alone [30] and we are dealing with two highly interactive streams, which requires us to consider the inter-stream dependence and temporal dynamics during topic extraction. Secondly, if there appears a new topic in UGC stream and it is mentioned in the subsequent news, we assume news is potentially driven by UGC and vice versa. The influence could be short-term (people talked about *Cruz* when his speech just ended), long-term (the discussion about *the budget bill* lasted throughout the event) or none, and we need to detect and distinguish different types of influence, as well as quantify the mutual influence. Thirdly, news is responsible for dealing with the controversial (or influential) topics in UGC, and we need to figure out what is an appropriate response from the processed events, such as how many topics the news respond to, how fast the response is and whether the public accept it.

To tackle the issues above, we introduce a three-phase framework: we first propose a novel cross dependence temporal topic model (CDTTM) to organize news and UGC into two dependent topic streams. The core idea of CDTTM is employing the dependent and temporal correlation across streams for building up two correlated generative processes with mutual reinforcement [16,30]. Then we develop a hybrid method to build links among the topics across streams based on KL-divergence and dynamic time wraping, which can effectively distinguish short or long-term influence. Finally, we systematically propose four statistical measures to quantify the mutual influence between two streams. Specifically, we introduce *topic progressiveness* to determine whether UGC goes ahead of news in some topics, *response rate*, *response promptness*, and *response effect* to evaluate how the news responds to UGC. Our main contributions include:

- We propose and formalize a novel problem to measure the mutual influence across two text streams and address it through a three-phase framework.
- We introduce a novel CDTTM model for topic extraction from two correlated text streams, which utilizes both temporal dynamics and mutual dependence.
- We propose a hybrid topic linking method, which can effectively discover the short-term and long-term influence links across streams.
- We define four metrics to quantify the influence between news and UGC streams. Experiments on five real news datasets show that the influence is bidirectional, namely news can trigger the generation of UGC and vice versa.

The rest of the paper is organized as follows. In Section 2, we formally define the problem of news and UGC influence analysis, and then demonstrate our methods. Our experimental results are reported in Section 3. Section 4 reviews the related literatures, and finally Section 5 concludes this paper with future research directions.

2 Problem and the Proposed Method

In this section, we formally define the problem of analyzing the influence between news and UGC streams, and then present our methods on topic extraction, link discovery and influence quantification.

2.1 Preliminaries and Problem Definition

Whenever an important ***event*** happens, it often brings forth two correlated text streams, namely news from media forms a news stream NS and users' voice from different social applications converges into a UGC stream $UGCS$. Each ***news*** d_i or user ***post*** p_i is represented by a content-time pair $(\mathbf{w_i}, t_i)$, where $\mathbf{w_i}$ denotes the words from a vocabulary V and t_i is the time stamp from a time collection T. Meanwhile, they both talk about several ***topics*** z_d or z_p, and the topics themselves keep changing along the timeline.

Definition 1. ***News and UGC Influence Analysis.*** *Given news stream NS and UGC stream $UGCS$, our goal is to extract time-ordered topic sets for both streams, namely Z_n and Z_u, discover the influence link collection $L = \{l_i = (z_x, z_y, \zeta)\}$ and characterize the mutual influence as several well-designed measures over the influence links. Note that $z_x \in Z_n$ and $z_y \in Z_u$ are topics from different streams NS and $UGCS$ respectively and ζ is a real number stating the link strength.*

Fig. 1 gives the NYT news and comment streams about *U.S. Federal Government Shutdown*, both talking about common topics like *budget bill*, *vote* and *debt ceiling*. The influence analysis aims to detect the dynamic topics for each stream, link the topics across streams, and evaluate how they influence each other.

According to the definition above, we present the solution in three phases: 1) ***topic extraction***: identify the topics from two correlated streams. 2) ***influence link discovery***: it is required to consider the influence types (i.e. short-term or long-term) when linking the common topics across streams. 3) ***influence quantification***: measure the intrinsic relations between UGC and news.

2.2 Topic Extraction from Two Text Streams

In this section, we extract the topics Z_n and Z_u from news and UGC streams. We observe that topics in these two streams are *cross-dependent*: comments in $UGCS$ are typically formed by the public to reflect their opinions on the topics published in news or provide new information about the progress of events, while the subsequent news reports in NS often provides additional information or clarifications to respond to the public comments in $UGCS$. To capture the temporal and cross-dependent information across streams, we expand the document comment topic model in [16] and design the cross-dependence temporal topic model (CDTTM) in Fig. 2.

We first introduce the notations. θ_d and θ_p are topic distributions for news and UGC and ϕ denotes the word distribution of each topic; x is a binary variable indicating whether the generation of the current word is influenced by the previous news (or UGC) ($x = 1$) or not ($x = 0$); α, β are the Dirichlet hyper parameters; λ is the *Bernoulli* parameter for sampling x and γ_d, γ_p are its hyper parameters.

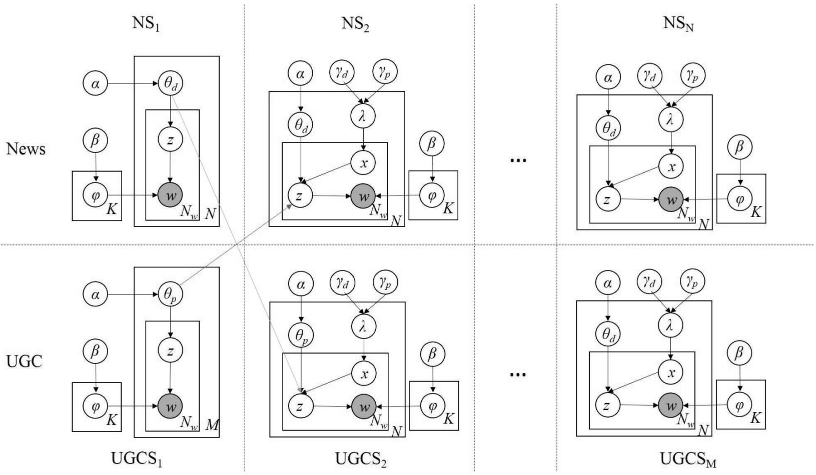

Fig. 2. Cross-Dependence Temporal Topic Model

For the generative process, we partition both steams into disjoint substreams with fixed time intervals, e.g. $NS = NS_1 \cup \ldots \cup NS_N$ where NS_i is with time

$[t_i, t_{i+1})$, and initialize two standard LDA models in the first substreams NS_1 and $UGCS_1$. As for the subsequent substreams, if there is a previous substream in the other stream (e.g. NS_{i+1} has a previous UGC substream $UGCS_i$), a coin x is tossed according to $p(x|d) \sim beta(\gamma_d, \gamma_p)$ to decide whether w_d inherits from the previous substream, otherwise (namely when there is no previous substream in the other stream) a standard LDA model is employed for word sampling. For example, we are currently sampling the word *majority* in news substream, and it appears frequently in the topic *budget bill* in previous UGC substream, which could serve as prior knowledge, namely, it has a higher probability to be assigned as topic *budget bill* in the current news substream as well.

For parameter estimation, we take Gibbs sampling technique for its ease of implementation. Suppose the previous UGC model is given, we sample the coin x and topic assignment of word w in the current news substream separately. For x, we derive the posterior probability:

$$p(x_i = 0|\mathbf{x}_{\neg i}, \mathbf{z}, \cdot) = \frac{n_{dx_0}^{\neg di} + \gamma_d}{n_{dx_0}^{\neg di} + n_{dx_1}^{\neg di} + \gamma_d + \gamma_p} \times \frac{n_{z_{di}}^{\neg di} + \alpha}{\sum_z (n_z^{\neg di} + \alpha)} \tag{1}$$

where n_{dx_0}, $n_{z_{di}}$ are the number of times that coin $x = 0$ and topic z has been sampled from d, and $\neg$ means exclusion. Then the posterior probability of topic z for word w_{di} when the coin $x = 1$ is derived as follows:

$$p(z_{w_{di}} = j|x_{w_{di}} = 1, \mathbf{z}_{\neg w_{di}}, \cdot) = \frac{n_{jw_{di}}^{\neg i} + m_{jw_{di}} + \beta}{\sum_w (n_{jw}^{\neg i} + m_{jw} + \beta)} \frac{n_{dj}^{\neg i} + m_{dj} + \alpha}{\sum_z (n_{dz}^{\neg i} + m_{dz} + \alpha)} \tag{2}$$

where $m_{jw_{di}}$ denotes the times that word w_{di} has been generated by topic j in current news substream and previous UGC substream respectively, $n_{dj}^{\neg di}$ and m_{dj} are the times that topic j has been sampled independently or influenced by previous UGC substream.

Readers who are interested in the solution can refer [16,26] for details. Through topic modeling, we turn news and UGC streams into two correlated and time-ordered topic streams Z_n and Z_u, where $Z_n^i \subset Z_n$ is the news topic set in time interval $[t_i, t_{i+1})$.

2.3 Topic Influence Link Discovery

In this section, we present our method to link topics across streams. Normally, we are dealing with three types of influence links, short-term, long-term and no influence.

Link Measurement. To measure if there is influence link between topics across streams, we calculate their distance using Kullback-Leibler(KL) divergence. Given two topics z_1 and z_2 associated with two distributions ϕ_1 and ϕ_2, the influence distance from z_1 to z_2 is defined as the additional new information in z_2 compared to z_1:

$$\zeta_{z_1 \rightarrow z_2} = KL(z_2||z_1) = \sum_{i=1}^{\mathcal{V}} \phi_{2i} \times \log \frac{\phi_{2i}}{\phi_{1i}} \tag{3}$$

where a larger $\zeta_{z_1 \rightarrow z_2}$ value indicates a weaker influence from topic z_1 to z_2. Note that the KL-divergence is asymmetric, but it makes sense in our scenario because the influence link strengths between two topics are not equal, i.e. news topics usually have higher influence to UGC topics than the other way around.

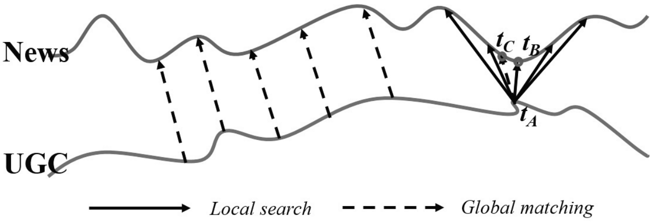

Fig. 3. Local search and global matching methods

Link Discovery. To accurately discover the influence between topics across streams, we perform both local and global search. Considering the short-term response between news and UGC, we set a time window $[-2, +2]$ for each topic in one stream to detect its local influence links, e.g. given a topic z in UGC stream, we search the topics locally that are most likely to influence z in the current and its previous two news substreams, and the topic that it may influence in next two news substreams. The linked topic is denoted as z_s representing the short-term influence. In Fig. 3, we link the topic in UGC stream at time t_A to its nearest point t_B in news stream within ±2 time intervals. To find the topic z_l with long-term influence, we first calculate the topic hotness along time for each topic, then employ dynamic time wraping (DTW) due to its high efficiency on dealing with time series data [11]. DTW is a class of widely-used dynamic algorithms which take two time series data as input, stretch or compress them in order to make one resemble the other as much as possible. As shown in Fig. 3, although t_B is the most appropriate point locally, we still link t_A and t_C and store it into z_l after making global matching on these two topic series.

Link Filtering. For each topic z at a specified time, z_s presents us a microcosmic view of the influence between topics within a pre-defined time interval while z_l discovers the macroscopic view of the influence. We compare z_s and z_l, and take the one with smaller KL-divergence as the final result. Finally, we sort all the linked pairs by distances and keep pairs whose distances are lower than the median value to remove those noisy or no influence cases, which constitute the result influence link collection L.

2.4 Influence Quantification

In this section, we infer the influence between news and UGC streams by defining four metrics in terms of news communication [21]. On the UGC side, we evaluate whether UGC influences news by introducing a novel concept of topic progressiveness, while on the news side, we evaluate how news reacts to UGC topics through adapting three popular measures in public opinion analysis.

Definition 2. ***Topic Progressiveness*** *tells whether UGC topics could trigger news topics. Inspired by [23], we consider the time difference between topics across streams. Specifically, for each topic z_p in UGC stream with linked topics $L(z_p)$ in the news stream, we compute the cross-entropy values to find the minimum one z_d^{min}, and then the topic progressiveness is defined as:*

$$Prog(z_p) = T(\arg\min_{z_d \in L(z_p)} H(z_p, z_d)) - T(z_p) \tag{4}$$

where $H(z_p, z_d)$ is the cross-entropy of z_p with its linked topic z_d, and $T(.)$ returns the time stamp of the input topic.

While it is clear and verified that news guides the generation of UGC [16, 30], the influence from UGC stream to news stream is our focus in this paper. Under this notation, if UGC topic z_p comes before news topic z_d^{min}, the topic progressiveness would be positive, meaning that z_p gives contribution to the z_d in news stream, and z_p is called as *progressive topic*. On the other hand a negative value of topic progressiveness indicates that news z_d^{min} has led to topic z_p in UGC stream.

Oftentimes, there emerges several hot or controversial topics in UGC stream, which might become public opinion crisis if they are not handled properly. Several models of cognitive determinants of social behaviors, e.g. the theory of planned behavior (TPB) [2,3], prove that intention is the most reliable predictor of behavior, but there is still a substantial gap between peoples intentions and their subsequent behavior due to many factors. Besides *perceived behavioral control*, the mediators play critical roles in the intention-behavior relation [27,28]. Since news is the most important mediator between the emergent topics and the public, we want to know whether and how much does news respond to the public topics, and if the public accept the response. Therefore, we quantify the following three metrics in public opinion management [9].

For convenience, we first classify the discovered links in previous steps into two groups. For a UGC topic z_p, if its linked news topic z_d appears earlier, we call z_d *previous linked topic* of z_p; otherwise it is *future linked topic* of z_p. The previous/future linked topics of z_d are denoted as $PL(z_p) = \{z_d | z_d \in L(z_p)$ and $T(z_d) < T(z_p)\}$ and $FL(z_p) = \{z_d | z_d \in L(z_p)$ and $T(z_d) > T(z_p)\}$.

Definition 3. ***News Response Rate (NRR)*** *defines how many UGC topics that draw news' attention. It is obtained through computing the percentage of topics that appear in UGC before news media:*

$$NRR(NS) = \frac{|\{z_p | FL(z_p) \neq \emptyset \&\& PL(z_p) = \emptyset\}|}{|\{z_p | PL(z_p) = \emptyset\}|} \tag{5}$$

Note that we only consider those topics without previous links, namely, they appear in UGC stream first.

Definition 4. ***News Response Promptness (NRP)****, which evaluates how fast news responds to the UGC topics, is calculated by the average time difference between the first appearance time in UGC and the response time in news:*

$$NRP(NS) = \frac{\sum_{z_p \ and \ z_d \in FL(z_p)} [T(z_d) - T(z_p)]}{\sum_{z_p} |FL(z_p)|} \tag{6}$$

Since z_d is *future linked topic* of z_p, $T(z_d) - T(z_p)$ should be positive consistently.

When a piece of news with topic z_p is published at a particular time t, it could address the voice and concerns in UGC stream via providing additional relevant information, and subsequently the public might answer whether they are satisified. In general, users often express their sentiments about certain topic using opinion words.

Definition 5. ***News Response Effect*** *is defined by checking if the number of opinion words has largely been reduced after news response. If so, it means that news has effectively address the concerns from the users.*

$$NRE(z_p, t) = \frac{C(z_p^{t-}) - C(z_p^{t+})}{C(z_p)} \tag{7}$$

where $C(z) = |\{w | w \in KW(z) \cap OP\}|$ *with a pre-defined opinion words set* OP*,* $KW(\cdot)$ *representing the keywords in given topics, and* t *the time stamp of the linked news topic.*

Remarks. *Progressiveness* presents a microcosmic view of the influence between topics across streams, while the other three metrics demonstrate the news response macroscopically. Particularly, larger *rate* and smaller *promptness* express that news responds to public topics actively and promptly, and large *effect* value indicates effective news response.

3 Experiments

In this section, we evaluate the proposed methods for topic extraction, influence link discovery and analysis. We first briefly introduce our datasets, and then present the detailed experimental results.

3.1 Data Preparation

To the best of our knowledge, no public existing benchmark data is available for analyzing the influence between news and UGC. Therefore, we have prepared five data of different events from influential news portals and social media platforms (e.g. NYT, Twitter), including *the Federal Government Shutdown* (cFGS/eFGS) in two languages, *Jang Sung-taek's* (Jang), *The Boston Marathon Booming* (Boston) and *India Election* (India). Particularly, we crawled news and comments about the first three events from specific pages[1,2] or through keyword search[3]. While for the last event from Twitter, we collected 2,890,801 related tweets using keyword filtering, then recognized all the tweets accompanied with URLs (there are 5,949 URLs found in 784,237 tweets), and finally news URLs whose frequencies are greater than 5 and corresponding tweets were selected.

For each dataset, we kept the continuous news reports and comments (or tweets), and further sorted them by published time (*last update time* for comments), and performed some cleaning work, such as removing low-frequency(≤ 3) words and stop words. The basic statistics after preprocessing are summarized in Table 1.

Table 1. Datasets: sources of two streams (Twitter news comes from various news websites, users post the shortened URLs along with their comments in tweets), duration, numbers of comments and news articles, max and average number of comments per news

Source	**Event**	**Days**	**Comments**	**News**	**Com./News**	
					max	avg
Sina-Weibo	cFGS	35	12,995	97	7,818	134
	Jang	43	3,291	84	467	39.2
NYT-Comment	eFGS	53	17,295	136	1,112	127
	Boston	46	7,521	211	518	29.4
News-Twitter	India	66	4,723	88	113	53.7

3.2 Topic Extraction

For the topic extraction process, we first compare the proposed method with several baseline models in perplexity, and then demonstrate the model result through a case study.

Perplexity Evaluation. To evaluate the topic model, we split the news and UGC by date to apply our proposed CDTTM model, and compare the results with the following methods:

- **DTM**: dynamic topic model which is proposed in [6]. We use the released version[4] on news and comments separately since it does not consider the cross dependence between the two streams.
- **DCT**: document comment topic model [16] which models a single news article and associated comments by considering the news-comment dependence. We implement this model, and adapt it to model multiple documents.

[1] http://news.sina.com.cn/zt/

[2] http://www.nytimes.com/pages/topics/index.html

[3] http://query.nytimes.com/search/sitesearch/

[4] https://code.google.com/p/princeton-statistical-learning/

- **TCM**: temporal collection model introduced in [15] which models temporal dynamics through associating the Dirichlet hyper parameter α with a predefined time-dependent function, and we implement the algorithm described in their paper.

We employ perplexity [7] of the held-out test data as our goodness-of-fit measure. A model with lower perplexity indicates it has good generalization performance.

As for the parameters, we set the number of topics $K = 5$, fix the hyper parameters $\alpha = 50/K$, $\beta = 0.1$, $\gamma_d = 5$, $\gamma_p = 0.2$ for news, and swap the values of γ_d and γ_p for UGC as recommended in [16,26]. Since we are modeling temporal data, we just perform the experiments by taking the prepositive several days for training and the rest for testing instead of random separation or cross validation. The results in Table 2 show that CDTTM performs better than the three state-of-the-arts. The reason for those with minor higher perplexity is that the test data is so sparse that CDTTM degenerates into standard LDA.

Table 2. Perplexity of four different topic models: the experiment settings include the duration, numbers of news articles and comments/tweets for both the *Train* and *Test* data, and the last four columns present the perplexity of different methods

Event	**Train**		**Test**		**Perplexity**			
	days	*docs*	*days*	*docs*	*DTM*	*DCT*	*TCM*	*CDTTM*
cFGS	28	74+10,175	7	23+2,820	19,717	19,204	18,071	17,923
Jang	10	71+1,979	33	13+1,312	17,203	16,211	17,146	17,307
eFGS	43	98+11,900	10	38+5,395	28,074	26,557	26,831	25,835
Boston	16	173+6,222	30	38+1,299	17,129	16,317	17,294	16,677
India	42	68+3,246	24	20+1,477	17,444	16,863	17,208	17,153

Case Study. Table 3 shows the results for eFGS. Note we have grouped the event into 5 stages according to their topic similarities and removed the common topic words (like *obama*) across multiple stages. We observe the topic trends from both streams (in the following description, **n** and **u** denote the information sources, i.e. news and UGC): [**n&u**]after the discussion (Aug. 29 to Sep. 19) on the Obamacare, [**n**]NYT published the news about the *Senate* and *House of representatives* discussing if they should support this program on Sep. 20. [**n**]Then *Ted Cruz* delivered an extremely long *speech* to argue that *Obamacare* was a disaster on Sep. 24. [**n&u**]They *voted*, *debated* and *voted* again (Sep. 30), [**n**]but the U.S. government still shut down on Oct. 1. During this period, interestingly, [**u**]the UGC from the public played crucial roles, e.g. they wanted a decision for the *budget* program (keywords *cost, debt*), and accelerated the vote (keyword *majority* on Sep. 25). [**u**]Public further appealed the *obamacare* should be *negotiated* and *passed* before the *shutdown*, which potentially led to [**n**]government re-opening on Oct. 17.

Table 3. An example for topic extraction on *eFGS* dataset: due to the space limitation, we list the top words with generative probability greater than 0.01 in each stage for both news and UGC, and those words that might link news and UGC are highlighted in bold

	8.29∼9.19	9.20∼9.25	9.26∼10.2	10.3∼10.17	10.18∼10.21
News	*health* .021	*health* .038	*health* .037	*shutdown* .021	*medicaid* .023
	debt .019	*senate* .021	*care* .022	***debt*** .020	*care* .019
	senate .017	***vote*** .017	*insurance* .012	*health* .014	***budget*** .014
	deficit .011	*congress* .012	***cruz*** .011	***debate*** .012	*national* .013
UGC	*care* .029	***cost*** .024	*gop* .026	***debt*** .032	*health* .025
	vote .016	*congress* .017	***obamacare*** .018	***believe*** .020	*care* .025
	job .015	***majority*** .017	***pass*** .017	*right* .015	***tax*** .024
	debt .014	*coverage* .012	***negotiate*** .016	***vote*** .013	*money* .011

3.3 Influence Link Discovery

We evaluate the link discovery from two aspects, namely the number and the correctness of the discovered links.

Table 4 shows the number of discovered links using different topic models on the eFGS data, and we can observe that the number of local links is much more than that of global links for all topic models indicating that most UGC influence to news is more timely than slowly; the introduction of (mutual) dependence benefits the link discovery and that's why other three methods tend to find more links than DTM.

Table 4. Number of discovered influence links on *eFGS* data, including the numbers of local/global links and the average links per day

	DTM	DCT	TCM	CDTTM
Local	77	89	94	97
Global	13	13	13	14
Average	1.698	1.925	2.019	2.094

To obtain the ground truth for correctness evaluation, we invite three annotators to build links (*Total*) between two topics series and we only include those links that at least two of them agree (*Agree*) in the following evaluation. We use *Hybrid* to denote our proposed method, and compare it with the simplified version that only uses local search (denoted as *Local*) as well as random linking (*Random*). Table 5 shows the annotated statistics and the comparison results. We can see that: the annotated agreement ratio is around 50% which indicates that it is really a tough task; the random linking quality is very poor, while both local search and hybrid method are 10 times better; the hybrid method can achieve comparable (even better) results to the human annotation.

Table 5. Link correctness comparison: number of all distinct annotated links (*Total*), number of links that at least two annotator agrees (*Agree*), and the performance of different strategies

Event	**Annotation**		**Comparison in F1**		
	Total	*Agree*	*Random*	*Local*	*Hybrid*
cFGS	177	84	4.27%	43.4%	46.4%
Jang	201	89	4.66%	42.0%	45.5%
eFGS	239	123	3.93%	46.4%	51.3%
Boston	213	117	4.08%	43.3%	45.6%
India	291	124	4.11%	43.6%	44.9%

3.4 Influence Quantification

Fig. 4 shows how the four measures change over time in different events (to reflect the general trends, we normalize all time spans into [0,1]). For the topic progressiveness, we can see that: 1) UGC indeed has guidance to the news report throughout event life cycles, and the influence mainly falls into the beginning part and decreases along the timeline. 2) The highest values often come from the key points of events (e.g. the shutdown day in eFGS). The reason why there are two peaks in Jang is that Kim Jong-un gave a speech on Jan. 1 which captured many user-concerned topics.

Then we evaluate the news response to the topics in UGC stream, and have the following observations:

Macroscopically. 1) The response rate increases initially and then decreases with time whereas the promptness follows an opposite trends. This indicates that UGC topics attract attention from the news (response rate) and news responds to them rapidly (response promptness), especially when the events are still *hot.* 2) For the response effect, at the beginning, the public increasingly use sentimental words to express their opinions. When the event reaches around halfway stages, the sentimental words have largely reduced, indicating that news is effective in responding to public opinions.

Microscopically. Very interestingly, we also observe that English news can respond more and faster than Chinese news for the event of Federal Government Shutdown, and the corresponding effect is more significant. A possible explanation is that it's a U.S. internal event and American people have a better understanding the gists of the event. This observation also tells us that the metrics themselves are correlated with each other, e.g. a larger progressiveness often leads to higher response rate and lower response promptness, and the consequent effect are more likely to be significant.

4 Related Work

Our work in this paper is related to several lines of research in text mining and streaming data process.

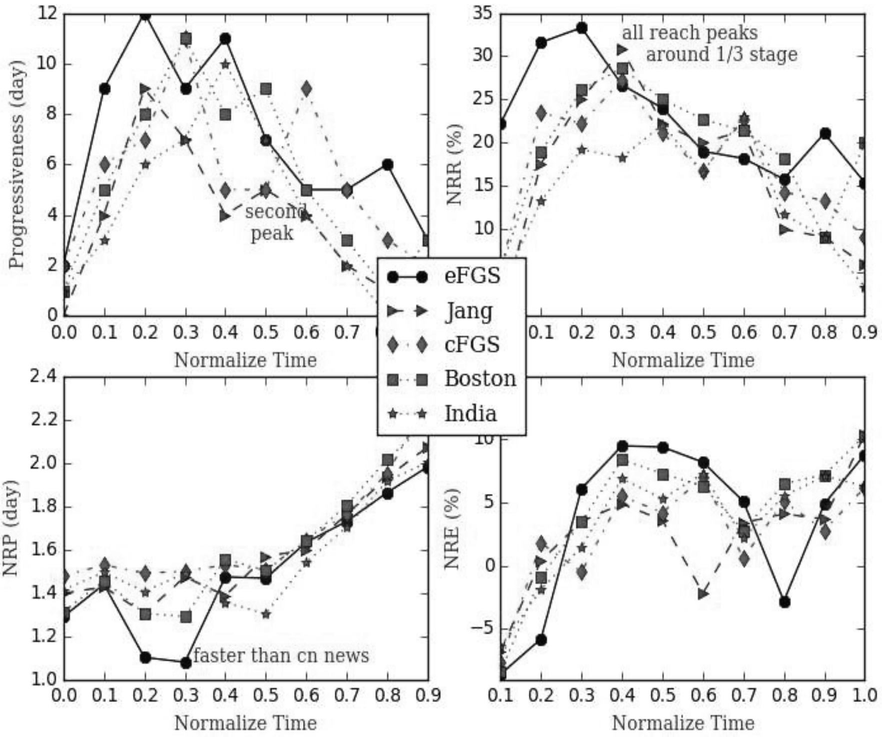

Fig. 4. Results Analysis for Influence Quantification

4.1 News and UGC Analysis

The rapid development of social media encourages many researchers to study its relationship between traditional news media. For example, Zhao et al. employed Twitter-LDA to compare the topic coverage of Twitter and NYT news and found Twitter actively helped spread news of important world events although it showed low interests in them [32]. Petrovic et al. examined the relation between Twitter and Newsfeeds and concluded that neither streams consistently lead the other to major events [25]. Liu et al. adapted the *Granger* causality to model the temporal dependence from large-scale time series data [5,8].

In this paper, we study the interplay of news and UGC in specific events where there should be more interactions than the general streams in the work mentioned above.

4.2 Streaming and Temporal Topic Model

For single stream model, Blei et al. proposed dynamic topic model to analyze the time evolution of topics in large document collections [6]. Wang et al. presented a non-Markov topic model named d Topic over Time (TOT), which jointly modeled both word co-occurrences and localization in continuous time [31]. Alsumait et al. proposed online-LDA to identify emerging topics and their changes over time in text streams [4]. Recently, Gao et al. derived topics in news stream incrementally using hierarchical dirichlet process, and then connected them using splitting and merging patterns [10].

For topic extraction in multiple streams, Wang et al. tried to extract common topics from multiple asynchronous text streams [29]. Hong et al. focused on analyzing multiple correlated text streams, allowing them to share common features and preserve their own private topics. Hou et al. proposed DCT model, which employed news as kind of prior to guide the generation of users' comments [16].

Compared with these models, the advantage of our model is that it captures the temporal dynamics and the mutual dependence between news and UGC streams.

4.3 Topic Evolution and Lifecycle

Mei et al. discovered evolutionary theme patterns in single text stream [22]. Wang et al. aimed at finding the burst topics from coordinated text streams based on their proposed coordinated mixture model [30]. Hu et al. modeled the topic variations through time and identifies the topic breakpoints [17] in news stream. Lin et al. formalized the evolution of an arbitrary topic and its latent diffusion paths in social community as an joint inference problem, and solved it through a mixture model (for text generation) and a Gaussian Markov Random Field (for user-level social influence) [20]. Jo et al. further captured the rich topology of topic evolution and built a global evolution map over the given corpus [18].

In this paper, we pay little attention on topic evolution and lifecycle, but more on analyzing the influence between two correlated text streams and try to figure out how news and UGC co-evolve along the time.

5 Conclusion and Future Work

In this paper, we study the mutual influence between news and UGC streams through a three-phase framework: extract topics from two correlated text streams, employ a hybrid method to discover short-term and long-term influence links across streams, introduce four metrics to measure the influence from UGC to news, as well as investigate how the news responds to the public opinion in UGC stream. Experiments on five news datasets confirm the existence of mutual influence, and present some interesting patterns.

There are several interesting directions to further extend this work. For example, our topic model returns a flat structure of topics and the topic number is pre-defined; it would be interesting to explore the hierarchical methods and non-parameter methods [10]. In addition, we only discover the influence links without distinguishing their different effects (e.g promote or suppress), so we will investigate the deeper semantics on the influence links which is another challenging but interesting problem.

Acknowledgments. Thank anonymous reviewers for their valuable suggestions that help us improve the quality of the paper. Thanks Prof. Chua Tat-Seng from National University of Singapore for discussion. The work is supported by 973 Program (No. 2014CB340504), NSFC-ANR (No. 61261130588), Tsinghua University Initiative Scientific Research Program (No. 20131089256) and THU-NUS NExT Co-Lab.

References

1. Ahmed, A., Xing, E.P.: 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. Ajzen, I.: From intentions to actions: a theory of planned behavior. Springer, Heidelberg (1985)
3. Ajzen, I.: The theory of planned behavior. Organizational behavior and human decision processes **50**(2), 179–211 (1991)
4. AlSumait, L., Barbará, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 8th IEEE International Conference on Data Mining, pp. 3–12 (2008)
5. Arnold, A., Liu, Y., Abe, N.: Temporal causal modeling with graphical granger methods. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 66–75 (2007)
6. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 113–120 (2006)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research, 993–1022 (2003)
8. 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)
9. Daily, C.Y.: Chinese monthly public opinion index. Tech. rep, China Youth Daily, December 2013
10. Gao, Z., Song, Y., Liu, S., Wang, H., Wei, H., Chen, Y., Cui, W.: Tracking and connecting topics via incremental hierarchical dirichlet processes. In: Proceedings of the 11th IEEE International Conference on Data Mining, pp. 1056–1061 (2011)
11. Giorgino, T.: Computing and visualizing dynamic time warping alignments in r: The dtw package. Journal of Statistical Software **31**(7), 1–24 (2009)
12. 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)
13. Hänska-Ahy, M.: Social media & journalism: reporting the world through user generated content. Journal of Audience and Reception Studies **10**(1), 436–439 (2013)
14. Holcomb, J., Gottfried, J., Mitchell, A.: News use across social media platforms. Tech. rep., Pew Research Center, November 2013
15. 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)
16. Hou, L., Li, J., Li, X., 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)

17. 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)
18. Jo, Y., Hopcroft, J.E., Lagoze, C.: The web of topics: discovering the topology of topic evolution in a corpus. In: Proceedings of the 20th International World Wide Web Conference, pp. 257–266 (2011)
19. Jönsson, A.M., Örnebring, H.: User-generated content and the news: Empowerment of citizens or interactive illusion? Journalism Practice **5**(2), 127–144 (2011)
20. 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)
21. McCombs, M., Holbert, L., Kiousis, S., Wanta, W.: The news and public opinion: Media effects on civic life. Polity (2011)
22. 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)
23. Mizil, C.D.N., West, R., Jurafsky, D., Leskovec, J., Potts, C.: No country for old members: user lifecycle and linguistic change in online communities. In: Proceedings of the 22nd International World Wide Web Conference, pp. 307–318 (2013)
24. Morinaga, S., Yamanishi, K.: Tracking dynamics of topic trends using a finite mixture model. In: Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining, pp. 811–816 (2004)
25. 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 (Poster) (2013)
26. Rosen-Zvi, M., Griffiths, T.L., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 487–494 (2004)
27. Sheeran, P., Abraham, C.: Mediator of moderators: Temporal stability of intention and the intention-behavior relation. Personality and Social Psychology Bulletin **29**(2), 205–215 (2003)
28. Sheeran, P., Orbell, S., Trafimow, D.: Does the temporal stability of behavioral intentions moderate intention-behavior and past behavior-future behavior relations? Personality and Social Psychology Bulletin **25**(6), 724–734 (1999)
29. 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)
30. 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)
31. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining, pp. 424–433 (2006)
32. 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)