A

Measuring the Semantic Uncertainty of News Events for Evolution Potential Estimation

XIANGFENG LUO, Shanghai University
JUNYU XUAN, Shanghai University and University of Technology Sydney
JIE LU, University of Technology Sydney
GUANGQUAN ZHANG, University of Technology Sydney

The evolution potential estimation of news events can support the decision making of both corporations and governments. For example, a corporation could timely manage its crisis public relations if a negative news event about this corporation is known with large evolution potential in advance. However, existing state-of-the-art methods are mainly based on time series historical data, which are not suitable for the news events with limited historical data and bursty property. In this paper, we propose a purely content-based method to estimate the evolution potential of the news events. The proposed method considers a news event at a given time point as a system composed of different keywords, and the uncertainty of this system is defined and measured as the semantic uncertainty of this news event. At the same time, an uncertainty space is constructed with two extreme states: the most uncertain state and the most certain state. We believe that the semantic uncertainty has correlation with the content evolution of the news events, so it can be used to estimate the evolution potential of the news events. In order to verify the proposed method, we present detailed experiment setups and results measuring the correlation of the semantic uncertainty with the content change of news events using collected news events data. The results show that the correlation does exist and is stronger than the correlation of value from the time-series based method with the content change. Therefore, we can use the semantic uncertainty to estimate the evolution potential of news events.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Natural Language Processing**]: Text analysis; I.5.4 [**Applications**]: Text processing

General Terms: ALGORITHMS, HUMAN FACTORS, EXPERIMENTATION

Additional Key Words and Phrases: Information Search and Retrieval, Text Mining, Natural Language Processing, News Event, Semantic Analysis

1. INTRODUCTION

The definition of an event in WordNet[1] is that "something that happens at a given place and time". In this paper, the events refer to the news stories, and their occurrences will trigger the reports from journalists or news websites and discussions or comments of people. We call this kind of events as 'news events' [Xuan et al. 2015], such as *MH370*

[1]https://wordnet.princeton.edu/

Author's addresses: X. Luo and, School of Computer Engineering and Science, Shanghai University, 200072, China, (email: luoxf@shu.edu.cn); J. Xuan, School of Computer Engineering and Science, Shanghai University, 200072, China, and Centre for Quantum Computation and Intelligent Systems (QCIS), School of Software, Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia, (email: xuanjunyu@shu.edu.cn); J. Lu and G. Zhang, Centre for Quantum Computation and Intelligent Systems (QCIS), School of Software, Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia, (email: Jie.Lu@uts.edu.au; guangquan.Zhang@uts.edu.au).

DOI:http://dx.doi.org/10.1145/0000000.0000000

missing. A news event[2] normally has many factors: when, who, what, where [Yuan et al. 2013], which are very important for the event detection. However, we focus on the content evolution of news events that are known in advance in this paper. So the above factors are not considered in this paper except the webpages.

Normally, the content of a news event, which is expressed by its webpages (e.g. reports of journalists and discussions of peoples), will change with time due to its evolution. For example, after the occurrence of the event *MH370 missing*, there were plenty of webpages that talked about it and the main discussions were about the *Aviation Accident*. Later, the people's attention had been paid on the *The Captain Hijacking*. This change of people's focus and evolution of a news story are both reflected by the change of content of the webpages of this news event. The problem is how to estimate the content change potential of a news event in the future.

The ability to estimate this content change potential of news events can be used in a variety of settings. For example, 1) prompt crisis public relation management for the corporation [Hennig-Thurau et al. 2003; Michelle L. Roehm 2006]. If it is known that the event *Coco-cola Contains Chlorine* will be paid a lot of attention and with large discussions tomorrow, the Coco-cola Corporation could manage its crisis public relation in time to minimize its economic losses and brand reputation damage; 2) The policy-making of the relevant departments of governments. Among a large number of news events on the web in each day, the events with large evolution potential could be recommended to the relevant departments which could pay more attention to and timely response to them. If a news event about *food security* has the potential to arouse much attention of people tomorrow, the relative department should response to it as soon as possible in order to reduce the public fear.

When aiming to esimate a variable, it is natural to use the temporal history of this variable [Cho and Garcia-Molina 2003]. The Auto Regressive Moving Average (ARMA) model [Brockwell and Davis 2009] would be the most classical tool for this kind of problem. For our purpose, the variable is the change of news events' content, so the straightforward idea is to utilize the temporal patterns of this change to estimate the evolution potential of a news event. However, this straightforward idea suffers from some problems: 1) it cannot handle the dynamic nature of news events, especially with burst property[3] [Barabasi 2011; Kumar et al. 2003]; 2) the content of news events are overlooked.

In this paper, we propose a purely content-based method to estimate the content change potential of news events. A hypothesis that the further content change of a news event is correlated with its current semantic uncertainty (that will be introduced later with more details) is put forward and verified. This hypothesis is the basis of the proposed content-based method. We start by mining a keyword association link network (EKLN) from the webpages to semantically represent the content of a news event at each time point (e.g. a day in this paper). The semantic uncertainty, which is a measurement of the evolution potential of a news event at current time point, is defined and computed through the complex network structure analysis of the EKLN using Information Entropy. In order to verify the correlation hypothesis, we estimate the content change potentials of 99 news events (with about 609,000 webpages) at each time point in their corresponding life circles by computing the semantic uncertainties based on the content of these news events. At the same time, the real content changes

[2]Since the content part of news events are with consideration, each news event can be seen as a topic and different aspects/sub-events could be seen as subtopics. So we do not distinguish the 'news event' with 'topic' in this paper.

[3]The 'bursty' of a news event here is in terms of the sudden change of simple statistics of this news event, such as webpage number.

between each consecutive time point pair is evaluated by a graph similarity algorithm as the benchmark. Finally, the correlation between semantic uncertainties and real content changes is obtained. This experiment shows that the correlation between the values by semantic uncertainty and the values of real change does exist. Besides, we also compare the proposed method with time series models in terms of the correlation through replacing the semantic uncertainties by the generated values from time series models. Experiments show that the values from the proposed method is more correlated with the real content changes of news events than the values from the time series models. It means that the proposed method is better than time-series-based models on the task of estimating the evolution potential.

The main contributions of this paper are as follows:

(1) An Event Keyword Link Network is mined from the webpages of a news event at a time point, which could capture the semantics of this news event at this time point and then make it computable;
(2) The semantic uncertainty of news event is introduced, formally defined and evaluated through the complex network structure analysis of the Event Keyword Link Network;
(3) The correlation between the semantic uncertainty and the content change of news events is verified through the real-world news event data.

The rest of this paper is structured as follows. In Section 2, we review related work. Basic definitions and the problem of this work are given in Section 3. The uncertainty of news events and proposed idea to estimate the content change potential are discussed in Section 4. A set of experiments are conducted in Section 5 and Section 6 concludes this study and discuses some future works.

2. RELATED WORK

Most related research works of this paper belong to the one topic of Information Retrieval area: Topic Detection and Tracking (TDT) [Allan et al. 1998; Hong et al. 2011a]. The basic tasks of TDT are: 1) to detect the appearances of new topics from the corpus and 2) to track the detected topics from the document stream. When the document corpus is about news webpages, the topics are also be named as news events. In this section, we categorize these works as three classes according the tasks: detection, tracking and prediction.

According to [Allan et al. 1998], event detection is the problem of identifying stories in a news stream. Several clustering algorithms are adopted to resolve this issue, like matrix factorization [Vaca et al. 2014], topic models [He et al. 2010] and non-homogeneous Poisson process [Huang et al. 2014]. It is interesting that Hassan et al [Sayyadi and Raschid 2013] has transformed the clustering to a community detection problem using a co-occurrence keyword graph to represent texts. Due to the great pervasive usage and real-time nature [Sakaki et al. 2010] of the social network, it becomes the best or comes one of the best disseminate and discuss news. The detection methods have been proposed based on the social network, but the different factors are considered, like geographic information [Yuan et al. 2013], named entities in the texts [Vavliakis et al. 2013], word clusters in the texts [Pervin et al. 2013], users' anomaly behaviours [Takahashi et al. 2014]. A keyword network is adopted by [Zhou and Chen 2014] that is similar to our method, but they only use the community detection on the constructed keyword network to detect events. It is interesting that click-through data [Zhao et al. 2006] is also used for the event detection. Our work is to do the further uncertainty analysis about the network to predict the content change/evolution of news events. The basic research objective of these state-of-the-art works is different from proposed method. In this paper, we assume that the events are known in advance

and we have already obtained the webpages about these events. All these work can be considered as the pre-processing procedure for obtaining events.

With the detected topics in hand, the further task is the tracking: how the topics evolve with time. This task is to depict the evolution graph or to discover the evolving patterns for the topics to assist people to understand these topics by the visualization methods. When constructing the evolution graph, the most important part is to define the relations between two temporal topics, such as the cosine similarity and Kullback-Leibler divergence [Mei and Zhai 2005] between the word vectors, the citation relation between member documents [Jo et al. 2011] and a relevance [Ha-Thuc et al. 2009]. When tracking topics, some side information is also incorporated, such as the social community evolution [Lin et al. 2010]. Similar with the topic evolution graph, the document dependent graph is constructed by the temporal dependency relations between documents for the better understand of a corpus [Shaparenko and Joachims 2007]. When mining the evolving patterns, researchers focus on discovering the interest temporal patterns of topics, like 'heartbeat'-like pattern [Leskovec et al. 2009]. In [Kamath et al. 2013], not only the temporal patterns but also the spatial patterns are studied. Mining the interest patterns from massive temporal topic data, the methods mainly belong to time series clustering problem [Yang and Leskovec 2011]. To sum up, both evolution graph mining and evolving pattern mining are based on the existing data, and there is no prediction involved despite they are very useful for people to understand a corpus or evolution of topics. However, the proposed method in this paper is to predict the content change of topics in the future not to mine the evolution graph or pattern.

Based on the models with or without the Markov assumption, we category the exsiting topic evolution study into two kinds: the first kind is with Markov assumption, like the Dynamic Topic Model [Blei and Lafferty 2006]. More examples are based on the 'Google Trends', some research work tries to predict the possible future events [Cho and Varian 2009], and the communities of blogspace are considered as topics and their temporal dynamics are also detected as the topic evolution analysis [Kumar et al. 2003]. On the contrary, Continuous Time Dynamic Topic Models [Wang et al. 2012b] is a non-Markov method that detects the temporal patterns of each topic by adding another time variable. Some researches have also considered the human factor, like sentiments [Wang et al. 2013] and smartphone locations[Kelly et al. 2013]. These state-of-the-art works have been certified to be successful in both theory and practical applications, such as where a real-world task to track volume of terms is realized [Hong et al. 2011b]. However, their representation of a web event is mainly based on keyword vectors. Actually, there is more information hidden in the webpages [Wang et al. 2007] of web events and there is a better way to preserve their semantics [Wang et al. 2006]. At the same time, they all rely on time series data. If there is only limited data, the prediction no longer works.

The prediction is apparently the most significant and difficult part in topic evolution analysis. It is just the research area that this paper belongs to. There are two main strategies for the prediction: one is based on causality relation and the other is based on correlation. The world knowledge ontologies mined from Linked Data[4] are used as source to learn the event causality relations for predicting the forthcoming events [Radinsky and Horvitz 2013; Radinsky and Bennett 2013; Do et al. 2011]. Some social manners, which is a kind of causality relation learnt from a document stream, are used to predict the topic distribution of new documents[Wang et al. 2012a]. These methods are strongly rely on the background knowledge about the events so that the causality relations mined from it will be acceptable enough. The aim of them is to

[4]http://linkeddata.org/

Table I: Notations in this paper

Symbol	meaning in this paper
e	a news event
t	a time point
K_t^e	word set of news event e at time t
R_t^e	word relation set of news event e at time t
Ω_t^e	event word link network of news event e at time t
U_t^e	semantic uncertainty of event e at time t
w_k	weight of kth word
$w_{k,t}$	weight of kth word at time t
wp	a webpage
K^{wp}	all words of a webpage
$\Delta_{t,t+1}$	the content change of an event between time t and $t+1$

predict the forthcoming events by current events, which is different from the aim of this paper that is to predict the content change of a single event. Besides, there is little content analysis about the events involved in these methods, but our proposed method is purely based on the content analysis. Similar content-based prediction is adopted by [Radinsky and Bennett 2013] for webpage content change prediction not for news event content change prediction. Another widely adopted method is correlation analysis. For example, the correlation between volume of queries and the economic activity is used for predicting the subsequent data releases [Cho and Varian 2009]. The persistence of topics and the resonance of the content are correlated [Asur et al. 2011]. The correlation between financial event and micro-blogging activity is used for predict the stock prices [Ruiz et al. 2012]. The proposed method in this paper is similar to this kind of work but with different research aims: our method is using correlation between the semantic uncertainty with the content change of a news event to predict the future change of this news event.

These state-of-the-art works have been confirmed to be very successful in their own research objectives: topic detection and tracking. For the prediction task, most of them are based on the time-series data from which the evolutionary patterns are discovered. These prediction models do not work well if only limited data is available, thus an approach is needed that does not rely on the historical data but only relies on the content of news events at a specific time.

3. DEFINITIONS AND PROBLEM FORMALIZATION

In this section, the basic concepts that will be used throughout this paper are defined, and the ultimate goal of the paper is clearly stated. The notations are summarized in Table. I.

Definition 3.1 (*News Event* e). A news event e is a unique thing that occurs at some place and time, and attracts broad attention of journalists and people, which leads to plenty of webpages emerging to report and discus it on the web.

Our definition is slightly different from the one in [Allan et al. 1998]. Here, we highlight one point that there must be some webpages that emerge to report and discuss them on the web. Only in this way can we analyze the news events through the analysis of the content of their webpages. Note that the factors, i.e. time and location [Yuan et al. 2013], are not mentioned in our definition because the news events used in this paper are known in advance. Our goal is to do the content analysis and estimation

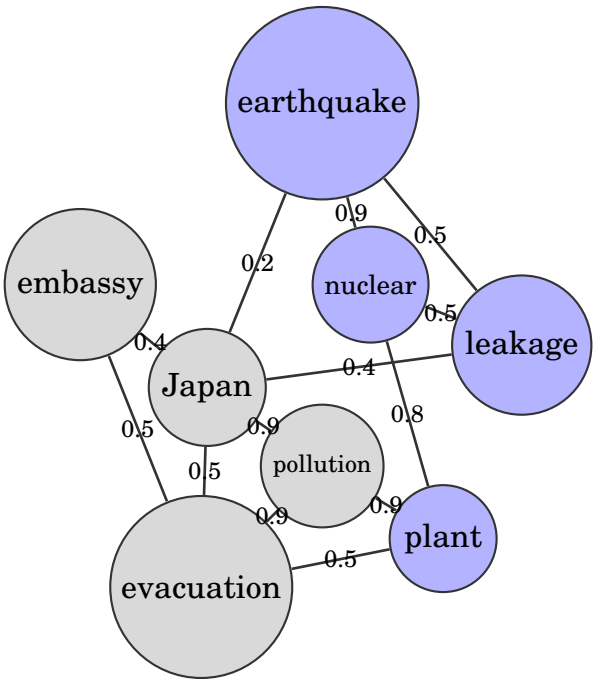

Fig. 1: An example of an Event Keyword Link Network (EKLN, Ω) of a news event, *Japan Nuclear Leakage*, on a given day. Each node of Ω represents a keyword of this news event, and the edge represents the association relation between two linked keywords (numbers on edges denote weights of relations). This EKLN will change along the evolution of this news event. Note that this is a simple man-made example for illustration only, and the real EKLN will contain more keywords and relations which can be automatically extracted from the webpages of this news event.

of the content change potential rather than event detection, so there is no need to explicitly provide these factors (they are considered as normal keywords here).

Since the webpages of a news event are unstructured data, we need a computable representation to do the further content/semantics analysis.

Definition 3.2 (*Event Keyword Link Network (EKLN)* Ω). An Event Keyword link Network of a news event e at time t, Ω_t^e, is a representation of this event's content/semantics at time t and composed of keywords and the association relations between these keywords, which can be formalized as,

$$\Omega_t^e = \{K_t^e, R_t^e\} \tag{1}$$

where K_t^e is the keywords set of a news event e at time t with weights and R_t^e is the relation set of e at time t with weights. The construction procedure is shown in Algorithm 1 and an example is given in Fig. 1.

This representation model is based on our former work: Association Linked Network (ALN) [Luo et al. 2011; Zhang et al. 2014]. ALN is a resource association model by a complex network structure [Zhang et al. 2014], which highlights the association relations between various resources and exhibits efficient performances on various applications [Luo et al. 2011; Zhang et al. 2014]. For EKLN, the nodes are keywords of a news event and the links are association relations between keywords. It should be

Algorithm 1 EKLN Construction

Input: A set of webpages of a news event $\{e\}$ at time point t
Output: EKLN Ω_t^e

1: Extract keywords $\{K_t^e\}$ from webpages by TF-IDF [Salton and Buckley 1988];
2: Extract the association rules $\{R_t^e\}$ between keywords from webpages by Apriori algorithm [Agrawal and Srikant 1994] using a webpage as a transaction;
3: Link the keywords $\{K_t^e\}$ using their association rules $\{R_t^e\}$ to form Ω_t^e.

Algorithm 2 Content Change Evaluation

Input: Two EKLNs of a news event at time point t and $t+1$: Ω_t and Ω_{t+1}
Output: Content Change $\Delta_{t,t+1}$

1: Compute node change by $\Delta^n_{t,t+1} = 1 - \frac{\sum_{k \in K_t^e \cap K_{t+1}^e} (|w_{k,t} - w_{k,t+1}|)/max(w_{k,t}, w_{k,t+1})}{|K_t^e \cup K_{t+1}^e|}$;
2: Compute edge change by $\Delta^r_{t,t+1} = 1 - \frac{\sum_{r \in R_t^e \cap R_{t+1}^e} (|w_{r,t} - w_{r,t+1}|)/max(w_{r,t}, w_{r,t+1})}{|R_t^e \cup R_{t+1}^e|}$;
3: Combine the node and edge changes together $\Delta_{t,t+1} = \frac{1}{2}(\Delta^n_{t,t+1} + \Delta^r_{t,t+1})$;

noted that EKLN is not just a simple combination of the keyword set and the association relation set; it is also a complex network which structure also has the ability to preserve the content/semantics of news events. As a complex network, EKLN has a bunch of network structures (e.g., small-world property [Zhang et al. 2014] and scale-free property) that can also be used to express the semantics or properties of a news event. These complex network structural properties cannot be directly observed from the association relation set but very important for our later uncertainty definition and evaluation.

There are indeed some alternatives for the representation of news events' content. One is Vector Space Model (VSM) [Salton et al. 1975] that represents a news event by a keyword vector. Compared with VSM, EKLN can preserve more content/semantics of news events. Another one is Web Ontology Language (OWL) [McGuinness et al. 2004] that represents a news event by some concepts and their concrete rich semantic relations normally predefined by experts. Compared with OWL [Wang et al. 2006; 2005], EKLN can be automatically constructed [Xuan et al. 2011] with the sacrifice of preserving less content/semantics of news events. For the news events on the web, it is really hard to construct ontology for them by experts considering their large scales and dynamic nature. Therefore, EKLN is an intermediate choice between the content/semantics preserving and automatic construction.

Definition 3.3 (*Evaluation Benchmark* Δ). The Evaluation Benchmark $\Delta_{t,t+1}$ of a news event e between time point t and $t+1$ is a measure of the difference of a news event's content at two time points. Considering the news event representation, the content change can be evaluated by the difference between two EKLNs at two time points.

The difference between two EKLNs is due to the difference of nodes and/or edges, which can be evaluated by graph matching algorithms [Gori et al. 2005; Zaslavskiy et al. 2009]. The one used in this paper is Vector Similarity Algorithm [Papadimitriou et al. 2010] as shown in Algorithm 2.

Definition 3.4 (*Semantic Uncertainty* U). The semantic uncertainty U_t^e of a news event e at time t is a measure of the evolution potential at a time point t.

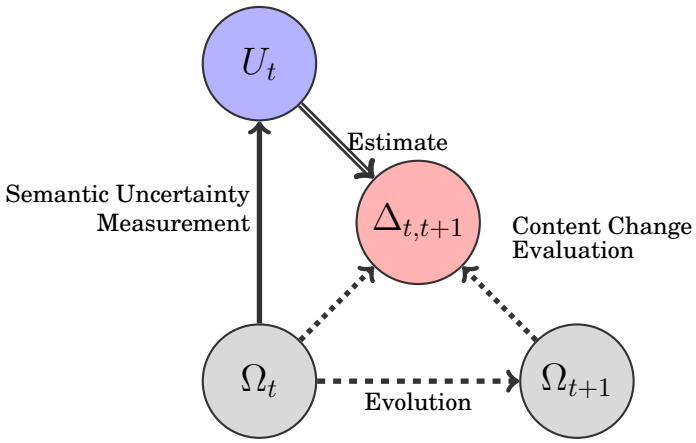

Fig. 2: The relations between Definitions in Section 3 and the illustration of the final aim of this paper. Ω_t is the EKLN of a news event at time t, and the U_t is the measurement of the semantic uncertainty of this news event Ω_t at time t. After the evolution, the news event has new state Ω_{t+1}, and there will be the content change $\Delta_{t,t+1}$ between Ω_t and Ω_{t+1}. The aim of this paper is to estimate $\Delta_{t,t+1}$ by U_t using the correlation between $\Delta_{t,t+1}$ and U_t.

What the semantic uncertainty means and how to evaluate it will be introduced in the following sections in detail.

The final aim of this paper is to use the current data of a news event to estimate its probability of further content change, as shown in Fig. 2.

The assumption that there are only webpages about a news event at a time point holds for two situations: 1) there is limited historical data about news events at the initial stage and 2) the former temporal patterns cannot help to estimate the future evolution because of the 'burst' property of news events. Both two situations provide little information for future evolution potential estimation, and what we have is only the webpages about a news event at a single time point.

We believe that the evolution potential of a news event is related to its current content. Our basic idea is to use the correlation between the semantic uncertainty of a news event to estimate its probability of content change.

4. SEMANTIC UNCERTAINTY OF NEWS EVENTS AND ITS MEASUREMENT

In this section, we introduce and measure the semantic uncertainty of news events, which, we believe, has correlation with the evolution potential of news events. At first, we re-weight the keywords according to their statistical and structural properties in the EKLN. Following that, we explain what the semantic uncertainty of news events is and how to measure it.

4.1. Keyword Weights in EKLN

Since keywords are the basic atoms for expressing the semantics of a news event, they are also the basis for the semantic uncertainty analysis for a news event. A better weighting scheme will lead to more accurate semantic uncertainty analysis. Therefore, we re-assign weights for keywords considering their 'positions' in a news event based on the original keyword weights from Algorithm 1. As introduced in Section 3, a news event at a given time is represented by an EKLN. The 'position' of a keyword in a news event denotes its weight in an EKLN. Here, we introduce two properties of keywords

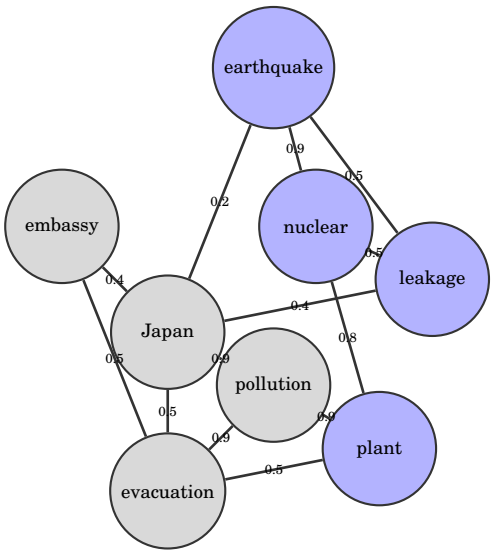

Fig. 3: An illustration of the keyword semantic expressing power of Document Frequency (DF). This EKLN has the same network structure with the one in Fig. 1, but the some semantics are apparently lost in this figure (all the nodes with same size).

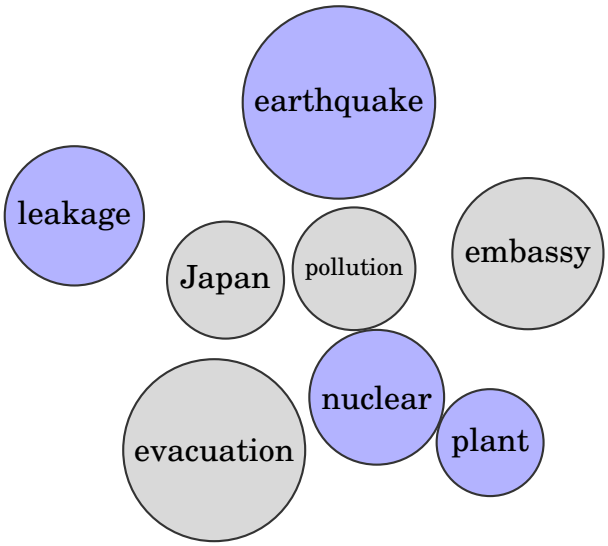

Fig. 4: An illustration of the keyword semantic expressing power of Network Structure (NS). Similar with the one in Fig. 1, all the nodes/keywords here have the same DF values, but the some semantics are apparently lost in this figure (the network structure is lost).

in an EKLN: statistical one and structural one, which will be used to re-assign the new weights to keywords.

Each node/keyword in an EKLN has a statistical property 'Document Frequency' (DF), which is the number of webpages containing this keyword,

$$w_k^{DF} = \sum_{all\ wp} \mathbf{1}(k \in K^{wp}) \tag{2}$$

where k denotes a keyword; wp denotes a webpage; K^{wp} is the keyword set of webpage wp; and the indicator function $\mathbf{1}(\cdot)$ equals 1 if wp contains k; 0, otherwise. The more frequently a keyword is mentioned in webpages of a news event, it is more likely that this keyword is the main semantic of this news event. (Note that the stopwords have already been removed.) For example, comparing Fig. 1, Fig. 3 loses some information because all keywords have the same DF which is represented by the same size of circles. We can easily identify that the keywords *earthquake* and *evacuation* are two dominant keywords of this news event at this time from Fig. 1 compared with Fig. 3. This information loss denotes the reduction of the semantic expressing power.

Another important property of nodes/keywords in an EKLN is 'Network Structure' (NS), which is an advantage of EKLN representation of news events comparing VSM representation. Comparing the Fig. 1, we can see that there is no organization of nodes/keywords in Fig. 4. Some information, i.e. the relation between *earthquake* and *leakage*, are lost, although the keywords have the same DF (sizes of circles) with the ones in Fig. 1. In order to capture this property by the keyword weights, we select Network Degree that is a basic metric for the nodes in a network. The Network Degree of a keyword is the number of other keywords connected with this keyword in an EKLN.

$$w_k^{NS} = \sum_{k' \in \Omega} neighbor(k, k') \tag{3}$$

where k and k' denote two keywords in Ω and $neighbor(x, y)$ is equal to 1 if keyword x and y are neighbors in Ω; 0, otherwise.

To sum up, the above discussions show that two different factors (Document Frequency and Network Structure) could impact the weights of keywords in a news event represented by the EKLN at a given time. Here, the following formula is taken to combine two properties together as the new weights of keywords in an EKLN,

$$w_k = \frac{1}{2}\left(\frac{w_k^{DF}}{\sum_k w_k^{DF}} + \frac{w_k^{NS}}{\sum_k w_k^{NS}}\right) \tag{4}$$

where w_k is the new weight of keyword k, $\sum_k w_k^{DF}$ is the sum of DFs of all keywords in an EKLN, and $\sum_k w_k^{NS}$ is the sum of Degrees of all keywords.

One doubt about the combination of DF and NS is that it will be meaningless if they have strong correlation with each other. In order to clarify this doubt, we have evaluated their correlations with real-world news event data. Their correlation coefficient is around 0.2, so we believe their combination is meaningful.

4.2. Semantic Uncertainty of News Events

As illustrated by the Fig. 5 and definition of EKLN, we know that the whole content/semantic of a news event at a given time is composed and expressed by different subtopics with their own semantics expressed by different keywords. The different keywords or different weights of the keywords will lead to different states of a news event. For example, assume that there are only two keywords used to describe a news event *Japan Nuclear Leakage*: *embassy* and *earthquake*, and they have the same weights in the EKLN of this news event. In fact, two keywords represent two subtopics of this event: *people evacuation of the embassy due to the nuclear radiation* and *the reason and impact of this earthquake*. When one subtopic has the dominate weight (i.e., the

weight of *embassy* is larger than *earthquake*), people can easily know that the content/semantics of this event is mainly about the *embassy*; when they have the same weight, it will be confusing for people which one is the dominate aspect of this event.

Here, a news event/EKLN can be considered as a system which is composed of keywords and their association relations. Different from the uncertainty of physical systems, the uncertainty of the news event system is from the keywords (the basic semantic atoms), so we call the uncertainty of news events as Semantic Uncertainty. Entropy has been successfully used to measure the uncertainty of a complex system. The entropy of the news event system is defined as,

Definition 4.1 (*EKLN Entropy*). The uncertainty of an EKLN can be measured by Entropy as,

$$H_\Omega = -\sum_{k=1}^{|K_t^e|} p_k \log p_k, \quad p_k = w_k / \sum_{k \in K_t^e} w_k \tag{5}$$

where H_Ω is the entropy value of Ω, $|K_t^e|$ is the keyword number in Ω, and p_k is the proportion of keyword k's weight to the sum of weights of all the keywords in Ω.

Through further analysis of the Eq. (5), we know that the EKLN entropy reaches its maximum value $\log |K_t^e|$ (the news event will be most uncertain) when all the keywords have the same weights,

$$p_1 = p_2 = \cdots = p_K \tag{6}$$

Therefore, we define the EKLN that satisfies this condition as the most uncertain state of a news event.

Definition 4.2 (*The Most Uncertain State of News Event,* $\widetilde{\Omega}$). The most uncertain state of a news event is the state that its all the keywords have the same weights in the corresponding EKLN.

However, EKLN entropy can only define the upper bound of the uncertainty of a news event at a given time. The value of the EKLN entropy could be infinitely close to zero, with the difference between the weights of keywords increasing. To construct an uncertainty state space for news events, an appropriate state should be selected as the lower bound of the uncertainty of news events. Next, we will discuss the lower bound of the uncertainty of news events based on power-law distribution.

The power-law distribution [Barabasi 2011][Adamic and Huberman 2000] has proven to be a common phenomenon in many areas, which is generally represented as,

$$f(x) \propto x^{-\alpha} \tag{7}$$

where x is the variable, $f(x)$ is the function of x, and α is the parameter. This distribution is similar with scale-free[5] and Zipf's[6], which all highlight that the the number of large x is less than the small x. In order to check if a variable satisfies power-law distribution, the log-log straight line fitting is normally adopted as

$$\log(f(x)) \approx -\alpha \log(x) + b \tag{8}$$

[5]http://en.wikipedia.org/wiki/Scale-free_network
[6]http://en.wikipedia.org/wiki/Zipf's_law

The error of log-log straight line fitting is used as the criteria of power-law distribution satisfaction,

$$error = \sum_{x} \frac{|\log(f(x)) + \alpha \log(x) - b|}{\sqrt{1+\alpha^2}} \tag{9}$$

Normally, when $error$ is smaller than a given value, x is deemed as satisfying power-law distribution. Here, we define a perfect power law distribution satisfaction in order to make a difference with the traditional satisfaction of the power law distribution,

Definition 4.3 (*Satisfaction of Perfect Power Law Distribution*). If the log-log curve of a distribution is and only is a rigid straight line ($error = 0$), this distribution is a perfect power law distribution.

Since EKLN is a complex keyword network, the most certain state of EKLN could refer to the most certain state of the complex network. So the problem is transferred to what is the most certain state of a complex network. In this study, we assume that EKLN has a tendency to reach power-law distribution, considering the social activity nature and the webpage corpus expression of news events. This assumption comes from the following literatures: (i) G. Bianconi [Bianconi 2008] suggests that, "The appearance of the power-law degree distribution reflects the tendency of social, technological, and especially biological networks toward 'ordering'. " A news event is essentially a social activity, so it may also have a tendency to be 'ordering'. In other words, EKLN as a representation of news events may also have a tendency to be power-law distribution; (ii) at the same time, Zipf's law (another expression of the power-law distribution) is observed in many languages and corpora. Many researchers have tried to explain the emergence of this phenomenon from different views [Piantadosi 2014]. Benoit Mandelbrot [Mandelbrot 1953] theoretically derives this distribution under the condition that it wants to minimize the average cost per unit of information (independent of languages). It means that when the keyword rank-weight distribution satisfies the Zipf's law, people can take the least effort to understand the content/semantics of a document corpus. For a news event, all the content/semantics are expressed by its webpages, so the keyword rank-weight distribution may satisfy the Zipf's law when the most certain state is reached. Note that there is no literature explicitly showing that a news event has a tendency to satisfy the power-law distribution. However, we made the above assumption based on the social activity nature and the webpage corpus expression of news events inspired by the existing literatures. Besides, this assumption will be used as the basis for the definition of the most certain state of news events and then used to measure the semantic uncertainty of the news event at a specific time. The experimental results in Section 5 have also empirically supported this assumption. We believe that this assumption could also be one contribution of this study.

The traditional 'power-law distribution' for a complex network [Barabási and Albert 1999] only considers the degrees of nodes in the network. But our power-law distribution of EKLN is the rank-weight distribution, which is

$$w_k \propto k^{-\alpha} \tag{10}$$

where all keywords are ordered by their weights w_k. According to the proposed assumption, we define the most certain state of the news events as the one that the keyword rank-weight distribution satisfies the perfect power-law distribution. Fig. 5 illustrates different degree of satisfaction of states of news events to the corresponding most certain states. A real example of the rank-weight distribution and corresponding fitting line is shown in Fig. 6.

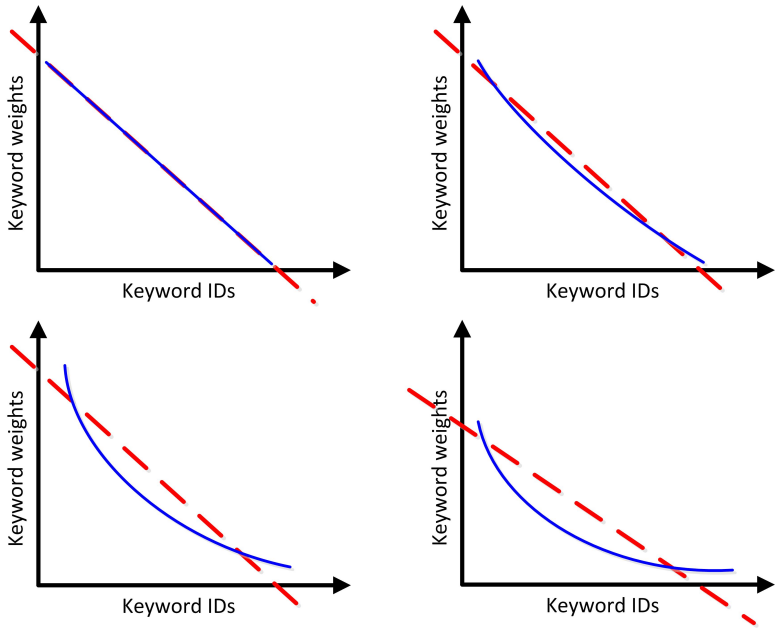

Fig. 5: The different satisfaction degrees of power-law distribution

Definition 4.4 (*The Most Certain State of a News Event,* $\widehat{\Omega}$). The most certain state of a news event is the state that the weights of all keywords in an EKLN satisfies the perfect power-law distribution.

After two extreme uncertainty states of news events have been defined: an upper bound from EKLN entropy and a lower bound from perfect power law distribution, a space of uncertainty states of a news events is fixed. Any news event at any time point will have its semantic uncertainty state within this space. Furthermore, the space of different news events or a news event at different time stamps will be different. For example, the most uncertain state has a relation with the number of keywords of a news event at a given time by the Eq. (5).

Following the discussion and definition of the semantic uncertainty, we next focus on the evaluation of the semantic uncertainty value. The semantic uncertainty space, bounded by the two extreme states: the most uncertain and certain states, has been constructed. Each news event at any time point has a corresponding state within this space. Referencing the defined two limited states, we evaluate the value of semantic uncertainty by the distances of current state to the limited states. So, the computation of semantic uncertainty is by,

$$U = \frac{|H_{\widetilde{\Omega_t}} - H_{\Omega_t}|}{d\left(\Omega_t, \widehat{\Omega_t}\right)} = \frac{|\log|K_t^e| - H_{\Omega_t}|}{\sum_k \frac{|\log(w_k)+\alpha\log(k)-b|}{\sqrt{1+\alpha^2}}} \tag{11}$$

where $|K_t^e|$ is the number of keywords in current EKLN and $H_{\widetilde{\Omega_t}}$ is the EKLN entropy of the most uncertain state. $|H_{\widetilde{\Omega_t}} - H_{\Omega_t}|$ is the distance from the current state to the most uncertain state. $d\left(\Omega_t, \widehat{\Omega_t}\right)$ is the distance from the current state to the most certain state. To sum up, the computation of Semantic Uncertainty is given in Algorithm 3.

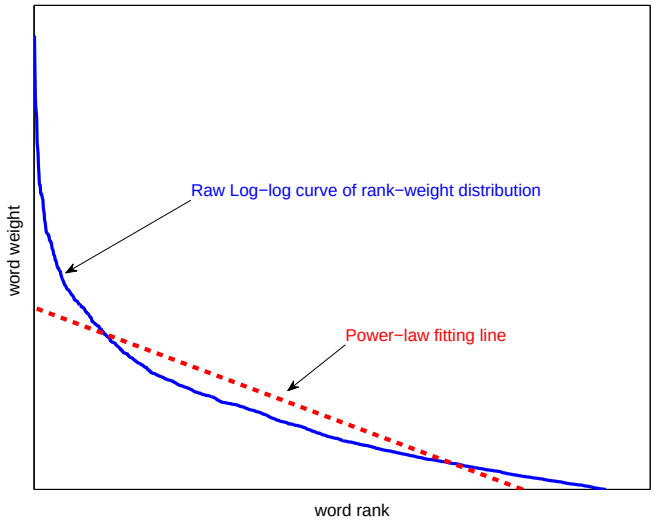

Fig. 6: A real example of the keyword rank-weight curve and its corresponding power-law fitting line. (News Event: *Fukushima explosions*; Date: Mar. 14, 2011; 1300 keywords)

Algorithm 3 Semantic Uncertainty Computation

Input: EKLN
Output: Semantic Uncertainty, U

1: Compute the keywords' weights $\{w_k\}$ according to Eq. (4);
2: Rank all the keywords by weight descending order to obtain the $\{w_k\}$ distribution curve;
3: Do the straight line fitting to the $\{w_k\}$ distribution curve, and then get the parameters α and b;
4: Compute H_{Ω_t} by Eq. (5);
5: Compute U by Eq. (11).

The Semantic Uncertainty is the measurement of the evolution potential of the news events. The larger the Semantic Uncertainty at time t is (i.e., the value of U is big), the more significantly the news event will evolve (i.e., the value of $\Delta_{t,t+1}$ is big). Our evolution potential estimation is just based on the correlation between the Semantic Uncertainty and the Content Change of news events. In the following section, we certify the correlation between the Semantic Uncertainty and the Content Change of news events through the collected real-world news events data.

5. EVALUATION AND DISCUSSION

In this section, we firstly introduce the dataset used in this paper. The second part is to describe the experiment setup for verifying the correlation between the semantic uncertainty of news events and their content change, followed by the discussions of the correlation results. Finally, comparative experiments are conducted to show the efficiency of the proposed method.

5.1. DataSet Collection

The webpages of news events used in this paper were collected from *Google*[7]. The collection procedure is as follows:

(1) Select a news event e, e.g. *MH370 missing*;
(2) Identify the first day t_s of this news event as the start time point by algorithm [Jin et al. 2010] with extra human involvement, and the end time is set as $t_e = t_s + 60$ except there is no webpages returned before this day;
(3) Start at $t = t_s$;
(4) Put the name of a news event e and the day t into the search engines;
(5) Rank all the returned webpages by relevance;
(6) Download webpages as much as possible;
(7) Redo Step (4) with $t = t + 1$; Until $t = t_e$.

We selected 99 hot events in China from March 2011 to May 2012 shown in *Baidu News* (http://news.baidu.com/)[8], because we started the data collection from May 2012. One selection criteria is that the event must last more than 10 days. This duration criterion is used to ensure that there are enough user discussions and reports on these events.

Note that the performance of EKLN depends on the effectiveness of extracted keywords although keyword extraction is not the contribution of this work. The EKLN is constructed as the semantic representation for a given news event. Therefore, it is naturally expected that all the keywords exist to express certain semantics of news events rather than 'noise' words, because the keywords (i.e., nodes in EKLN) are the basic semantic atoms. However, these keywords are from the webpages which are written using the natural language, so there must be some noise words within these webpages too. For example, some words, such as 'though', 'but' and 'am', are used only for the natural language expression but not for the event semantic expression. The main objective of keyword extraction algorithm is to filter out these words.

The EKLN has a capability to reduce the influence from the noise words from keyword extraction to the uncertainty computation to some extent. This capability is from the association relations between words. Some nodes may be the noise words from keyword extraction step but they normally have seldom association relations with other meaningful words, so these words will be posited at the edge of network and then their influences will be small to the uncertainty computation.

TF-IDF, as a renowned keyword extraction algorithm, could remove the words with too large document frequency, but it is sensitive to the collection of webpages because it is based on statistics. While removing the meaningless words (i.e., 'is' and 'but'), TF-IDF may remove some important words which are also with large document frequency. The accurate filtering of these words with large document frequency is very important, because the remaining ones will be at important positions (i.e., with large degree) in the EKLN. The words with small document frequency will tend to be posited at the edge of EKLN. The influence of the missing or redundant of words with large document frequency is significantly greater than the missing or redundant of words with small document frequency to an EKLN and then to the uncertainty computation of news events. Therefore, for EKLN, a perfect keyword extraction algorithm should accurately remove the meaningless ones but keep the meaningful ones especially for the words with relatively large document frequency.

[7] www.google.com.hk
[8] *Baidu News* is only used to select the names of hot web events, but their webpages are downloaded using *Google*.

Table II: Statistics of Dataset

WebPageNum	609,696
ComparingDaypointNum	3,880
WebEventNum	99
AvgDayNum of a WebEvent	40
Data Source	*http://www.google.com.hk*

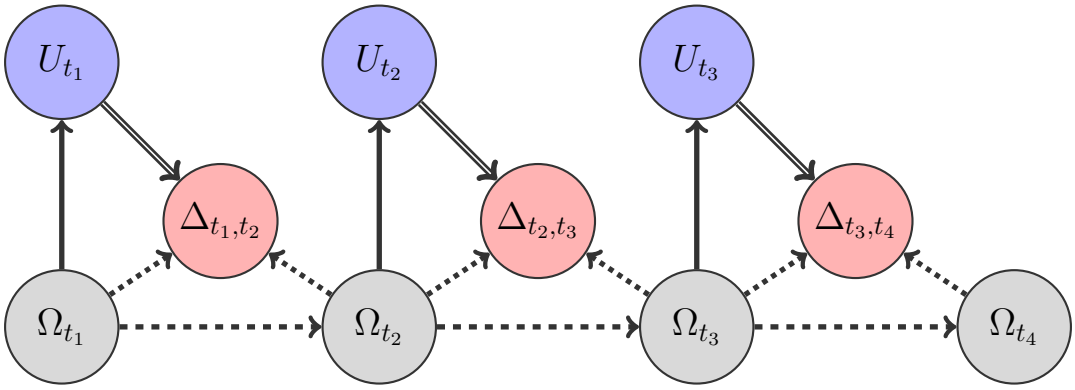

Fig. 7: An illustration example of the experiment setup. Suppose a news event has webpages at four time points. 1) firstly construct EKLN at each time point; 2) compute the semantic uncertainty at each time point; 3) evaluate the real content change between each consecutive time pair; 4) finally the correlation between U and Δ is evaluated.

The webpages of each news event were downloaded and preprocessed, including word segmentation, stopword removal, POS tagging, keyword extraction and the association relations of keywords mining. Some statistics are shown in Table II. The names of these news events are listed in Table IV of Appendix.

5.2. Experiment Setup

With the collected news event dataset in hand, we design a method to evaluate the correlation between the Semantic Uncertainty and the Content Change in this section.

The experiment setup is designed as: Considering a news event e with only two time points, we compute U_1 based on its EKLN Ω_1^e at the first time point. At the same time, we evaluate the Content Change $\Delta_{1,2}$ between this event at first and second time points. The value of $\Delta_{1,2}$ is used as a benchmark here, because it is computed by the real data. Then, we can evaluate the correlation between U_1 and $\Delta_{1,2}$. The strong correlation means that U_1 can be used to estimate $\Delta_{1,2}$. An illustration example is shown in Fig. 7.

For a news event e with T time points, we firstly construct EKLN at each time point,

$$\overrightarrow{\Omega^e} =< \Omega_1^e, \Omega_2^e, \ldots, \Omega_T^e >$$

Following that, a content change benchmark for this news event e is obtained by evaluating the Content Change $\overrightarrow{\Delta}$ for each consecutive time point pair,

$$\overrightarrow{\Delta} =< \Delta_{1,2}, \Delta_{2,3}, \ldots, \Delta_{T-1,T} > \tag{12}$$

Table III: Statistics of all events between Semantic Uncertainty and Content Change in Fig. 9.

	≥ 0	≥ 0.3	≥ 0.5
Number	92	67	40
Percentage	92.93%	67.68%	40.40%

With the benchmark in hand, we can evaluate the correlation. The Semantic Uncertainty $\overrightarrow{U}$ at each time point (except the last time point) is computed,

$$\overrightarrow{U} =< U_1, U_2, \ldots, U_{T-1} > \tag{13}$$

Note that both $\overrightarrow{U}$ and $\overrightarrow{\Delta}$ are vectors with T-1 dimensions. Finally, we can compute their correlation by Pearson Correlation Coefficient ($coco$) and the equation is,

$$coco = \frac{\sum_{t=1}^{T-1}(U_t - \overline{U})(\Delta_t - \overline{\Delta})}{\sqrt{\sum_{t=1}^{T-1}(U_t - \overline{U})^2}\sqrt{\sum_{t=1}^{T-1}(\Delta_i - \overline{\Delta})^2}} \tag{14}$$

where $\overline{\Delta}$ is the mean of $\overrightarrow{\Delta}$ and $\overline{U}$ is the mean of $\overrightarrow{U}$. Then, for each news event, we can obtain a Correlation Coefficient.

It will be a little confusing that if we have the data about a news event at each time point, why cannot we just use the Algorithm 2 to evaluate the Content Change $\Delta_{t,t+1}$ and then the Semantic Uncertainty is useless? The reason is that in the real-world application setting, we do not have the future data Ω^e_{t+1}. What we have is only the data at current time, so we only have Ω^e_t. At this situation, we can only obtain the Semantic Uncertainty based on Ω^e_t but cannot obtain $\Delta_{t,t+1}$.

In the experiment, what we want to obtain is the correlation between the Semantic Uncertainty and the Content Change. Although the whole life cycle of each news event has already been collected as data, the computation of Semantic Uncertainty only uses the data at one time point. The Content Changes $\Delta_{t,t+1}$ evaluated from the data of news event are used as benchmark for comparison.

5.3. Experimental Results

Six news events are shown in Fig. 8 as examples. Each subfigure denotes a news event in which the x-axis denotes time points (day) and the y-axis denotes the value of Content Change and Semantic Uncertainty. There are two curves in each subfigure: the red (dash) curve denotes Content Change benchmark in Eq. (12) and the blue curve denotes the Semantic Uncertainties in Eq. (13). The correlation coefficient is labeled on the top of each subfigure. Since the different news events may have different life cycles, the length of x-axis of all subfigures are different.

The experimental results on all the news events are shown in Fig. 9, and the statistics related to Fig. 9 are shown in TABLE III. In Fig. 9, there are two subfigures: the top one shows the correlation coefficients of all the news events and the bottom one shows the p-values of corresponding correlation coefficients. It should be noted that the ordering of news event IDs in two subfigures are the same. P-value[9] is used to show the significance of the correlation coefficients. The smaller (often ≤ 0.05) the p-value is, the more significant the corresponding correlation coefficient is and the more confidence of the correlation conclusion. The average of correlation coefficients is 0.3870. It can be seen that the percentage of the number of news events ≥ 0.3 from F_I is 67.68%, which can be compared with the random value 35% (a random value of $[-1, 1]$ that is

[9]http://en.wikipedia.org/wiki/P-value

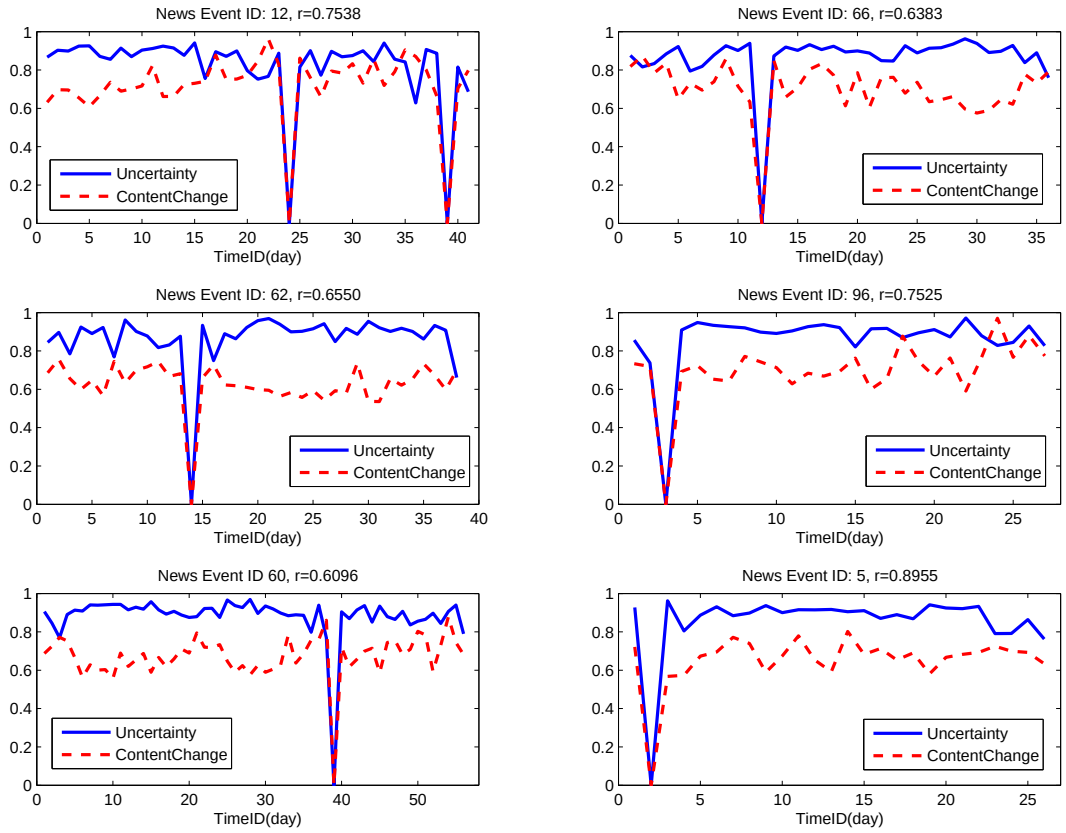

Fig. 8: The value of Semantic Uncertainty and Content Change of six selected news events on each day. The x-axis denotes the Time/Day ID. The p-values of all these news events are smaller than 0.05.

bigger than or equal to 0.3). Similarly, the percentage of the number of ≥ 0.5 from F_I is 40.40%, which can be compared with the random value 25% (a random value of $[-1, 1]$ that is bigger than or equal to 0.5). These comparisons show the non-random property of the value of Semantic Uncertainty U. We can draw the conclusion that there does exist the correlation between the Semantic Uncertainty and Content Change of news events.

More interestingly, we have found that the p-values of the big correlation coefficients are normally statistically significant (smaller than 0.05), but the ones of the small correlation coefficients are not (larger than 0.05). There are 32 news events with statistically insignificant p-values and three of them with correlation values larger than 0.5, while there are 67 news events with statistically significant p-values and four of them with correlation values smaller than 0.5. Based on these correlations, we can draw two conclusions with the results in Fig. 9 in hand: 1) It is *sure* that the values from our proposed method have significant correlation with the benchmark on the events (with *high* correlations and *significant* p-values in Fig. 9); 2) It is *not sure* that the values

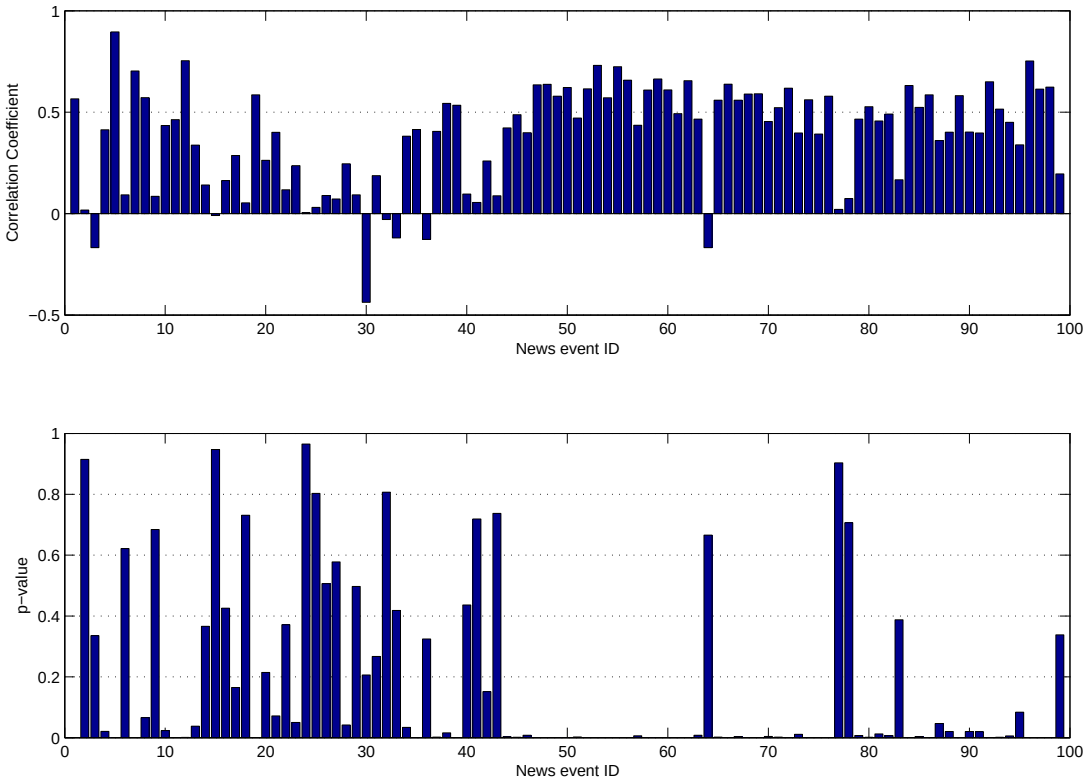

Fig. 9: Correlation Coefficients of Semantic Uncertainty with Content Change of all the news events. The top subfigure: the x-axis is the news event IDs (the order has no meaning), and the y-axis is the correlation coefficient of CR with F_I of a given news event. The bottom subfigure: the x-axis is also the news event IDs (the order is same with the top subfigure), and the y-axis is the p-values of the corresponding Correlation Coefficients in the top subfigure.

from our proposed method have a correlation with the benchmark on the events (with *small* correlations and *insignificant* p-values in Fig. 9). See the second conclusion again. We cannot draw the conclusion like this: the values from our proposed method do not have significant correlation with the benchmark on the events (with small correlations and insignificant p-values in Fig. 9), because the p-value is not significant for these events. From conclusion 1), the remaining news events with large correlations support our idea. From conclusion 2), we know that the events with small correlations are not against our idea.

However, the correlation is not very strong, because the evolution of news events may be also influenced by other factors, i.e., the social environment, which is not discussed in this paper. Social environment here denotes the former events in the same environment (i.e., a country) with the current event. To explore the influence of the social environment, we manually select two groups of news events: one with news events that are strongly influenced by the social environment and the other with news events that are weakly influenced by the social environment. If the correlation coefficients of

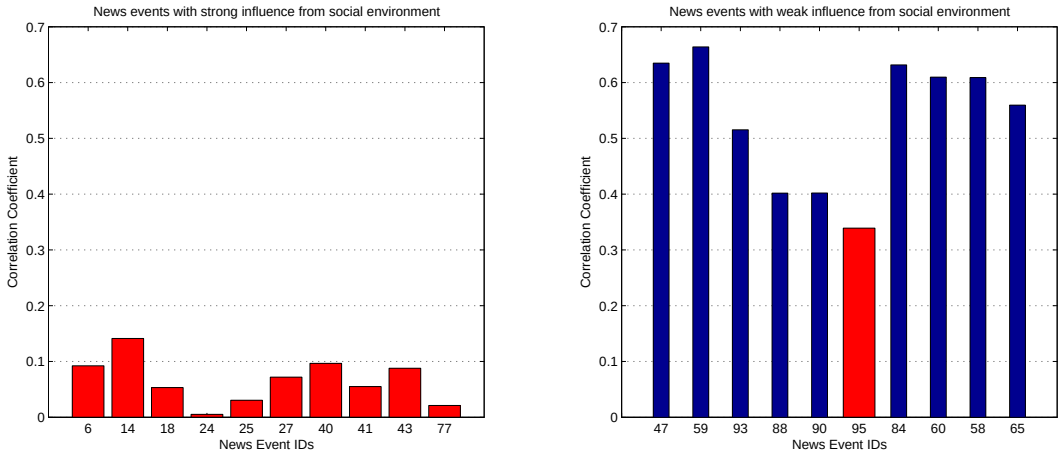

Fig. 10: Two group of news events with different influences from social environment: the left subfigure shows news events with strong influence from social environment and the right subfigure shows news events with weak influence from social environment. The x-axis denotes the news event IDs. The red (wide) bars denote the correlation coefficients with big p-values (bigger than 0.05), and the blue (narrow) bars denote the correlation coefficients with small p-values (smaller than 0.05).

the first group are lower than those of the second group, it means that the reason why the overall correlation coefficients are relative on some news events is due to the influence of the social environment. In Fig. 10, the values of the correlation coefficients of two groups are shown; the left subfigure of Fig. 10 is for the first group (under strong influence of the social environment) and the right subfigure of Fig. 10 is for the second group (under weak influence of the social environment). It is clear that the first group has small values of correlation coefficients. The evolutions of the news events in this group are influenced ont only by the semantic uncertainty but also significantly by their social environments. By contrast, the second group has very high correlation coefficient values (average 0.5366), which means that the intervention of the social environment is too small to influence the evolution of news events comparing to semantic uncertainty. It is interesting that we have found that the p-values of the correlation coefficients of the second group are small (smaller than 0.05) except the news event 95 (its p-value is 0.08 and also the smallest correlation coefficient in this group), but the p-values of the correlation coefficients of the first group are not smaller than 0.05. It means that the correlation coefficients of the second group is more significant and confident than the ones of the first group. Therefore, we can draw the conclusion that there does exist strong correlation between the Semantic Uncertainty and Content Change of news events, irrespective of the influence of social environment.

5.4. Comparative Experiment

In order to show the efficiency of the proposed method, we do an experiment to compare the performances of our method with a time series-based method and a regression method. Here, AutoRegressive Moving Average model (ARMA)[10], which is one of the most basic and important models for time-series analysis, is adopted, This model can

[10]http://en.wikipedia.org/wiki/Autoregressive%E2%80%93moving-average model

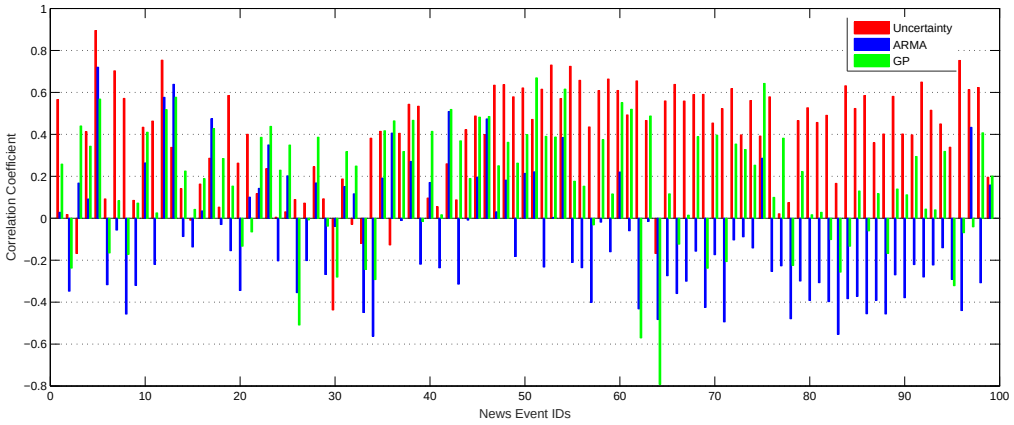

Fig. 11: Correlation Coefficients of Semantic Uncertainty with Content Change and predication by ARMA and GP with Content Change of all the news events. The x-axis denotes the news event IDs. The y-axis denotes the correlation coefficient.

be formalized as,

$$X_t = \mu_t + \sum_{i=1}^{p} \varphi_i X_{t-1} + \sum_{j=0}^{q} \theta_j \epsilon_{t-j} \tag{15}$$

where p and q are orders, ϵ is white noise and θ and φ are the parameters. More details can be found in [Brockwell and Davis 2009]. The implementation in Matlab is adopted here.

Another comparative method is Gaussian Process (GP) regression model [Williams and Rasmussen 2006],

$$f = GP(mf, \Sigma_f) \tag{16}$$

where f is a function, mf is a mean function, and Σ_f is a kernel function. GP is reported as a state-of-the-art regression model in machine learning area. The different selections of mean and kernel functions will have different hyper-parameters. Here, we use a linear mean function and Matern covariance function. The more details can be found in [Williams and Rasmussen 2006]. Its implementation[11] is used in this paper.

For a news event e, ARMA and GP estimate its content change,

$$< \Delta(\Omega_1^e, \Omega_2^e), \ldots, \Delta(\Omega_{t-1}^e, \Omega_t^e) > \xrightarrow{estimate} \Delta(\Omega_t^e, \Omega_{t+1}^e) \tag{17}$$

Comparing the proposed method, ARMA and GP need the former change time-series information $[1, t]$ but the proposed method is only based on data at one former time t. The results are shown in Fig. 11 and Fig. 12.

The Fig. 11 shows the final correlations between the content changes/evolutions of 99 news events with the generated scores by the proposed method, ARMA, and GP. The red bars are from the proposed method (same with Fig. 9), the blue ones are from ARMA, and the green ones are from GP. We can see that the ratio of (positive and high) correlations from proposed method is bigger than the one from ARMA and GP. We also compare the p-values of these correlations, as shown in Fig. 12. The three pies denote

[11] http://www.gaussianprocess.org/gpml/code/matlab/doc/

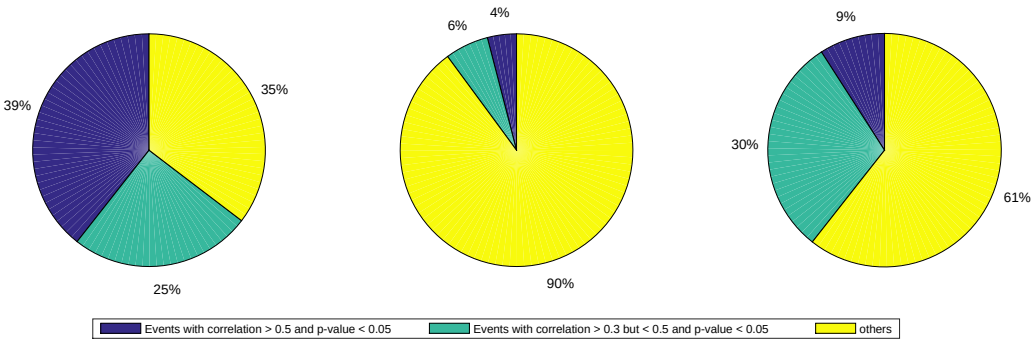

Fig. 12: Correlation Coefficients with p-values from three methods: Semantic Uncertainty, ARMA, and GP. The left pie chart is from Semantic Uncertainty; the middle one is from ARMA; the right one is from GP. In each pie chart, the blue area denotes the percentage of news events with correlation bigger than 0.5 and p-values smaller than 0.05; the green area denotes the percentage of news events with correlation bigger than 0.3 and p-values smaller than 0.05; the yellow area is the rest.

the results of three methods. In each pie chart, the blue area denotes the percentage of news events with correlation bigger than 0.5 and p-values smaller than 0.05; the green area denotes the percentage of news events with correlation bigger than 0.3 and p-values smaller than 0.05; the yellow area is the rest. It is apparent that the results from the proposed method achieve the best results. Consider the situation that the big correlations with statistically significant p-values and the small correlations also with statistically significant p-values. It means that both the big and small correlations have big confidences. Then, we cannot draw the conclusion that the generated score by the proposed method has correlation with the change/evolution of news events in this situation. However, it can also see from Fig. 12 that the small correlations tend to have statistically insignificant p-values from the proposed methods. It means that the big correlations with statistically significant p-values have big confidence and the small correlations with statistically insignificant p-values have small confidence. Therefore, we can say that the generated score by the proposed method has correlation with the change/evolution of news events.

6. CONCLUSIONS AND FUTURE STUDY

In this paper, we have defined the semantic uncertainty of news events and correlated it with the content change/evolution of these news events for the evolution potential estimation. The proposed method is a purely content-based method that resolves the issues of traditional time-series-based methods, i.e., limited historical data situation and the 'burst' property of the news events. The experimental results demonstrate that there does exist the correlation with the semantic uncertainty and the content evolution of news events. Furthermore, the comparative experiment with two classical time-series methods shows that the correlation between the semantic uncertainty with the content evolution is higher than the one between the generated value from the time-series method with the content evolution. Therefore, we can draw the following conclusions that: 1) there is correlation between the semantic uncertainty and the content change of news events; 2) using the semantic uncertainty to estimate the future content evolution potential of news events has better performance than time-series-based methods, i.e, ARMA and GP regression.

In the future, we will try to collect more news event data from different areas and countries to compare the proposed method within different domains. The influence of the social environment will also be an interesting research point. Another theoretical research will be the explanation of the relations between the parameters of the 'perfect power law distribution' of the EKLN with the evolution of news events.

ACKNOWLEDGMENTS

Research work reported in this paper was partly supported by by the National Science Foundation of China under grant no.61471232, and by the Australian Research Council (ARC) under discovery grant DP140101366 and the China Scholarship Council.

REFERENCES

Lada A Adamic and Bernardo A Huberman. 2000. Power-law distribution of the world wide web. *Science* 287, 5461 (2000), 2115–2115.

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*. 194–218.

Sitaram Asur, Bernardo A. Huberman, Gábor Szabó, and Chunyan Wang. 2011. Trends in social media: persistence and decay. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*.

Albert Barabasi. 2011. *Bursts: the hidden patterns behind everything we do*. Plume.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.

Ginestra Bianconi. 2008. Most probable degree distribution at fixed structural entropy. *Pramana* 70, 6 (2008), 1135–1142.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. ACM, New York, NY, USA, 113–120.

Peter J Brockwell and Richard A Davis. 2009. *Time series: theory and methods*. Springer.

Hyunyoung Cho and Hal Varian. 2009. *Predicting the present with Google trends*. Technical Report. Google Inc.

Junghoo Cho and Hector Garcia-Molina. 2003. Estimating frequency of change. *ACM Transactions on Internet Technology* 3, 3 (2003), 256–290.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 294–303.

Marco Gori, Marco Maggini, and Lorenzo arti. 2005. Exact and approximate graph matching using random walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 7 (July 2005), 1100–1111.

Viet Ha-Thuc, Yelena Mejova, Christopher Harris, and Padmini Srinivasan. 2009. A relevance-based topic model for news event tracking. In *Proceedings of the 32Nd International Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 764–765.

Qi He, Kuiyu Chang, Ee-Peng Lim, and Banerjee A. 2010. Keep it simple with time: a reexamination of probabilistic topic detection models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 10 (2010), 1795–1808.

Thorsten Hennig-Thurau, Gianfranco Walsh, and Gianfranco Walsh. 2003. Electronic word-of-mouth: motives for and consequences of reading customer articulations on the internet. *International Journal of Electronic Commerce* 8, 2 (2003), 51–74.

Dzung Hong, Qifan Wang, Dan Zhang, and Luo Si. 2011a. Query expansion and message-passing algorithms for TREC microblog track. In *TREC*. Citeseer.

Liangjie Hong, Dawei Yin, Jian Guo, and Brian D. Davison. 2011b. Tracking trends: incorporating term volume into temporal topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 484–492.

Shihang Huang, Ying Liu, and Depeng Dang. 2014. Burst topic discovery and trend tracing based on Storm. *Physica A: Statistical Mechanics and its Applications* 416, 0 (2014), 331–339.

Xin Jin, Scott Spangler, Rui Ma, and Jiawei Han. 2010. Topic initiator detection on the world wide web. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 481–490.

Yookyung Jo, John E. Hopcroft, and Carl Lagoze. 2011. The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 257–266.

Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. 2013. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 667–678.

D. Kelly, B. Smyth, and B. Caulfield. 2013. Uncovering measurements of social and demographic behavior from smartphone location data. *IEEE Transactions on Human-Machine Systems* 43, 2 (March 2013), 188–198.

Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. 2003. On the bursty evolution of blogspace. In *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*. ACM, New York, NY, USA, 568–576.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. ACM, New York, NY, USA, 497–506.

Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, and Jiawei Han. 2010. PET: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. ACM, New York, NY, USA, 929–938.

Xiangfeng Luo, Zheng Xu, Jie Yu, and Xue Chen. 2011. Building association link network for semantic link on web resources. *IEEE Transactions on Automation Science and Engineering* 8, 3 (July 2011), 482–494.

Benoit Mandelbrot. 1953. An informational theory of the statistical structure of language. *Communication Theory* 84 (1953).

Deborah L McGuinness, Frank Van Harmelen, and others. 2004. OWL web ontology language overview. *W3C Recommendation* 10, 2004-03 (2004), 10.

Qiaozhu Mei and ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*. ACM, New York, NY, USA, 198–207.

Alice M. Tybout Michelle L. Roehm. 2006. When will a brand scandal spill over, and how should competitors respond? *Journal of Marketing Research* 43, 3 (2006), 366–373.

Panagiotis Papadimitriou, Ali Dasdan, and Hector Garcia-Molina. 2010. Web graph similarity for anomaly detection. *Journal of Internet Services and Applications* 1, 1 (2010), 19–30.

Nargis Pervin, Fang Fang, Anindya Datta, Kaushik Dutta, and Debra Vandermeer. 2013. Fast, scalable, and context-sensitive detection of trending topics in microblog post streams. *ACM Transactions on Management Information Systems* 3, 4 (2013), 1–24.

StevenT. Piantadosi. 2014. Zipf word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review* 21, 5 (2014), 1112–1130.

Kira Radinsky and Paul N. Bennett. 2013. Predicting content change on the web. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 415–424.

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 255–264.

Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. 2012. Correlating financial time series with micro-blogging activity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 513–522.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 851–860.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513–523.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (1975), 613–620.

Hassan Sayyadi and Louiqa Raschid. 2013. A graph analytical approach for topic detection. *ACM Transactions on Internet Technology* 13, 2 (2013), 1–23.

Benyah Shaparenko and Thorsten Joachims. 2007. Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*. ACM, New York, NY, USA, 619–628.

Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi. 2014. Discovering emerging topics in social streams via link-anomaly detection. *IEEE Transactions on Knowledge and Data Engineering* 26, 1 (2014), 120–130.

Carmen K. Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. 2014. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, New York, NY, USA, 527–538.

Konstantinos N. Vavliakis, Andreas L. Symeonidis, and Pericles A. Mitkas. 2013. Event identification in web social media through named entity recognition and topic modeling. *Data and Knowledge Engineering* 88, 0 (2013), 1–24.

Chong Wang, David Blei, and David Heckerman. 2012b. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298* (2012).

Chao Wang, Jie Lu, and Guangquan Zhang. 2005. A semantic classification approach for online product reviews. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI '05)*. IEEE, 276–279.

Chao Wang, Jie Lu, and Guangquan Zhang. 2006. Integration of ontology data through learning instance matching. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '06)*. IEEE, 536–539.

Chao Wang, Jie Lu, and Guangquan Zhang. 2007. Mining key information of web pages: a method and its application. *Expert Systems with Applications* 33, 2 (2007), 425 – 433.

Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang. 2013. SentiView: sentiment analysis and visualization for internet popular topics. *IEEE Transactions on Human-Machine Systems* 43, 6 (Nov 2013), 620–630.

Yu Wang, Eugene Agichtein, and Michele Benzi. 2012a. TM-LDA: efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 123–131.

Christopher KI Williams and Carl Edward Rasmussen. 2006. Gaussian processes for machine learning. *MIT Press* 2, 3 (2006), 4.

Junyu Xuan, Xiangfeng Luo, Guangquan Zhang, Jie Lu, and Zheng Xu. 2015. Uncertainty analysis for the keyword system of web events. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* PP, 99 (2015), 1–1. DOI:http://dx.doi.org/10.1109/TSMC.2015.2470645

Junyu Xuan, Xiangfeng Luo, Shunxiang Zhang, Zheng Xu, Huimin Liu, and Feiyue Ye. 2011. Building hierarchical keyword level association link networks for web events semantic analysis. In *Proceedings of the 9th International Conference on Dependable, Autonomic and Secure Computing (DASC '11)*. IEEE, 987–994.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 177–186.

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, New York, NY, USA, 605–613.

Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. 2009. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 12 (Dec 2009), 2227–2242.

Shunxiang Zhang, Xiangfeng Luo, Junyu Xuan, Xue Chen, and Weimin Xu. 2014. Discovering small-world in association link networks for association learning. *World Wide Web* 17, 2 (March 2014), 229–254.

Qiankun Zhao, Tie-Yan Liu, Sourav S. Bhowmick, and Wei-Ying Ma. 2006. Event detection from evolution of click-through data. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. ACM, New York, NY, USA, 484–493.

Xiangmin Zhou and Lei Chen. 2014. Event detection over twitter social media streams. *The VLDB Journal* 23, 3 (June 2014), 381–400.

Appendix

Table IV: News events in Dateset
(The names are translated from their corresponding English names.)

No.	Name	No.	Name
1	Libya Unrest	**51**	living cost ranking
2	shouting font	**52**	LinShuHao best players
3	Shuanghui Clenbuterol	**53**	the US degree factories
4	Nuclear radiation panic buying	**54**	extracting live bear bile for medicine
5	Apple iPad2	**55**	Ipad trademark Weiguan
6	reserve ratio increase	**56**	Vanke formaldehyde exceeds bid
7	Zhangziyi Micrioblog	**57**	Laser correction sequelae
8	Supermoon	**58**	Express real-name
9	housing prices control target	**59**	Sanya cheating customer
10	Cosmetic raising prices	**60**	maldives coup
11	The international oil prices climb	**61**	Technology male female editor
12	Nuclear Radiation iodide pills	**62**	Angang spray explosion accidents
13	Fukushima explosions	**63**	Nanjing Nagoya break off
14	Hainan offshore duty-free	**64**	Syria conflict
15	food crisis	**65**	Girl disfigure case
16	Yaojiaxin	**66**	Hebei chemical plant explosion
17	Graveyard prices rose	**67**	BaiJing stabbed to death
18	'two-child' policy	**68**	Occupation Obama
19	delay retirement	**69**	Ipad trademark case
20	individual income tax threshold	**70**	Screw you tumor ruler
21	tainted steamed	**71**	Political Exam -1 score
22	Mengniu poisoning	**72**	Ipad3 publish
23	Vegetables price farmer	**73**	Girl Men's occupation
24	illegal food additives	**74**	Fangzhouzi Hanhan
25	World Earth Day	**75**	Relatives speech outline
26	flooded coal mine in Guizhou	**76**	Master's urban management
27	Apple Google tracking	**77**	National football team Jordan
28	power shortage	**78**	Adel captured
29	Forbes list's Chinese	**79**	workwoman leftovers fire
30	USA tornado	**80**	Film bureau fares
31	high-speed rail accident	**81**	recreational activities public funds
32	Elope microblog	**82**	relationship America north Korea
33	City life quality rankings	**83**	Jordan litigation
34	Shanxi leader PS	**84**	Push-ups death
35	Icel Volcanic Eruption	**85**	chief daughter dazzle rich
36	Taiwan plasticizer	**86**	Guojingming Forbes
37	Chilean volcano eruption	**87**	aboned baby safety isl
38	News of the World hacking	**88**	Air quality new national stard
39	Rice additives	**89**	Arab spring
40	Tsinghua university named	**90**	Hainan underground bank
41	Pork price	**91**	Baojie layoffs
42	The Red Cross GuoMeiMei	**92**	City water price
43	Da Vinci fraud	**93**	Shengda Literature IPO
44	Dalian petrochemical fire	**94**	Employment policy Account
45	Gansu school bus accident	**95**	LinZuLuan home village
46	Kim jong il's death	**96**	Russia's vote
47	Taobao mall Tmall	**97**	Political consultative conference
48	FengFeiFei lung cancer	**98**	National People's Congress
49	Whitney died	**99**	Congo serial explosion
50	grammy Adele		