World Wide Web
https://doi.org/10.1007/s11280-019-00753-2

Web event evolution trend prediction based on its computational social context

Junyu Xuan[1,2] · Xiangfeng Luo[1] · Jie Lu[2] · Guangquan Zhang[2]

Received: 13 April 2019 / Revised: 4 September 2019 / Accepted: 28 October 2019 /
Published online: 14 March 2020

Abstract
Predicting future trends of Web events can help significantly improve the quality of Web services, e.g., improving the user satisfaction of news websites. Existing approaches in this regard are based mainly on temporal patterns mined with the assumption that enough temporal data is available on hand. However, most Web events do not have a long lifecycle, but a burst property, which drastically reduces the performance of temporal patterns mining. Furthermore, these approaches overlook the influence of the social context surrounding the Web events. In this paper, we propose a novel method to predict future trends of Web events, based on their social contexts rather than temporal patterns. More specially, in the proposed method, a computational model for the social context is first built as a two-layer Association Linked Network considering its properties, such as the associative network property and the small world property. Then, the interaction between a Web event and the social context is simulated, based on the anchoring theory. Finally, an external force is defined and evaluated to quantify the influence of the social context on the evolution of Web events, which is used to predict future trends of Web events. Experiments show that the performance of the proposed method is better than that of the traditional time series-based approaches.

Keywords Web event · Evolution analysis · Social context · Human society · Human associative memory

1 Introduction

An event identifies something (nontrivial) happening in a certain place at a certain time [69]. Many events capture the public attention after their occurring, and among them, some events have been extensively reported and discussed on the Web, named Web events [39, 67],

This article belongs to the Topical Collection: *Computational Social Science as the Ultimate Web Intelligence*
Guest Editors: Xiaohui Tao, Juan D. Velasquez, Jiming Liu, and Ning Zhong

✉ Xiangfeng Luo
luoxf@shu.edu.cn

Extended author information available on the last page of the article.

such as *Birds Flu* and *Bushfire Sydney*. The content of a Web event (i.e., the reports and discussions of the public about this event on the Web) may change/evolve over time, which refers to 'event evolution'[1] [10, 31, 40]. For example, *News International Phone Hacking Scandal in U.K. 2011* can be regarded as a Web event mapped by its real corresponding event in the British society. At the very beginning of this event, the discussions and media coverage centered mainly on *phone hacking*. As this event developed, the focus of cyber citizens shifted to *personal privacy*. Following that, the *David Cameron news corporation* captured the attention of the public. At the event's final stage, *corruption* became the focus of discussions. The Web event progressed from *phone hacking* to *corruption*, because the issue of *corruption* is so sensitive that it attracts the attention of many people in most countries. From this example, it is evident that (1) the content of a Web event changes/evolves with time; (2) the changes/evolutions of Web events are influenced by the prevailing social context. For a Web event, its social context is constituted by the past Web events that can be seen as the expression of public interests and concerns in that social context (The formal definition of this is given in Section IV). Note that the evolution of Web events is different from their propagation: the propagation models [35, 70] aim to understand and predict how a Web event (news or rumors) propagates from one carrier (e.g., a website or a person) to another carrier; our evolution model aims to predict the change of public interests or attentions through analyzing the content of webpages.

The ability to predict 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 [23, 44]. If it is known that the event *Coco-cola Contains Chlorine* would attract the public attention, the Coco-cola Corporation could manage its crisis public relation in time to minimize its economic losses and brand reputation damages; 2) The policy-making of relevant departments of the government, and 3) predict the user click on popular news events [25, 57]. 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 on 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.

A Web event, in different social contexts,[2] may evolve in different ways. For example, assumed a Web event, called *New York Giants win Super Bowl*. Discussions about this Web event in *USA* are more than the ones in *China*. This is because, in *China* very few people are interested in this game, unlike in the *USA*, where many people like this game. Their discussions will contribute a lot of new information to that event and thus make the Web event evolve. Even in the same social context, two different Web events may evolve in different ways. For example, *Americans rocket launch* and *North Korea rocket launch* took place in the same international social context. The discussions about *Americans launch rocket* centred on *satellite usage* and *international space cooperation*, whereas those about *North Korea rocket launch* centred on *missile test* and *force threat*. The reason for this contrast is that the two Web events, though evolved in the same social context, can evoke discussions on different topics. To sum up, the contents of both the social context and Web events can affect their mutual interaction, and thus the evolution of Web events too. How to evaluate this influence is what constitutes the main objective of this paper.

[1] The evolution here does not mean that one event changes/eolves to another one, but the content of webpages following this event change/evolve over time.

[2] The social context here mainly denotes the context on the Web, but we want to keep this name to highlight its social nature.

Existing works on Web event (topic) evolution prediction rely on the data available for a period of time. Temporal text corpus mining algorithms and time series-based approaches are the main techniques to discover the historic temporal evolution patterns in data. But, they do not work well with Web events having 'burst' property [4, 30, 68], and (or) those having limited historical data. This is because the existing models or algorithms rely heavily on the historical temporal patterns, but both 'burst' and 'limited historical data' can reduce the efficiency of temporal patterns mining.

Our innovative idea is to utilize the social context and the content of a Web event at a specific time, rather than temporal patterns, to predict the Web event evolution. In this paper, we propose a computational social context model to semantically gather past Web events, considering their properties, such as the weights of past Web events, hot degrees, relationships with current Web events, activation degrees etc. Our objective thus resolves into addressing three tasks: (i) how to reasonably model the social context to make it computable, (ii) how to model the interaction with a Web event, and (iii) how to quantify the influence of social context on the evolution of Web event. During the modelling, we borrow the results in Sociology and Psychology to bring our model closer to reality, in terms of the interaction between the Web events and the social context. The experiments show that, even without historical temporal patterns, we can well predict the evolution of Web events based on our proposed social context-based method.

The following are the three major contributions of this paper:

1. A social context model is built based on Association Linked Network (ALN) to make the social context computable;
2. The interaction between the social context and its Web event is simulated through associated relations between keywords on both sides according to human memory theory, which facilitates quantification of the influence of the social context on the evolution of the Web event;
3. Design an equation to compute external force as the quantified of influence of social context on the evolution of a Web event, based on the social context model and the interaction simulation.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 formalizes the problem; Section 4 models the social context and Section 5 quantifies social context's influence; Section 6 develops the evaluation metric, besides providing the experimental results and Section 7 summarizes the findings and conclusions of this study.

2 Related work

In this section, we present the research findings relating to the analysis of Web event evolution topics and to social environment models.

The evolution analysis of topics/events focuses mainly on constructing their evolving topology or route map. For example, the dataset contains themes, which are split according to different time stamps [43]. However, it is found that the topology of topics can be discovered without splitting the data into time slots [27], and the topology or the evolving map so discovered is very useful. Some 'world knowledge' [52] or causality relations between events [51] also are incorporated into the storylines during their construction stage.

Also, some prediction models are proposed in addition to the construction of evolving topology, which forms the main research theme of this paper. Two main strategies are available for prediction: one is based on causality relation and the other on correlation. For predicting forthcoming events, the world knowledge ontologies, mined from LinkedData, are used as a source to learn the event causality relations [16, 50, 52]. These methods relied heavily on the background knowledge about the events, so that the causality relations mined from it will be readily acceptable. Their aim is to predict forthcoming events by current events, which is different from the aim of this paper, which is predicting the content change of a single event. Besides, in these methods, there is hardly any content analysis of the events, whereas the method proposed by us here is based purely on content analysis. Similar content-based prediction is adopted by [50] for webpage content change prediction, not for Web event content change prediction. Another widely adopted method is correlation analysis [71]. For example, the correlation between volume of queries and the economic activity is used for predicting subsequent data releases [13]. The correlation between financial event and micro-blogging activity is used for predicting stock prices [53]. The method proposed in this paper is similar to this kind of work, but differs in its research aims: it uses correlation between semantic uncertainty and content change of a Web event to predict future changes of this Web event.

Under the Markov assumption, the prediction models can be classified into two categories. Most of the temporal predictions, such as the Dynamic Topic Model [7], are under the Markov assumption. More examples are based on the ‘Google Trends’, some research work tries to predict the possible future events [13], and the communities of blogspace are considered as topics and their temporal dynamics are also detected as the topic evolution analysis [30]. In contrast, Continuous Time Dynamic Topic Model [61] is a non-Markov method that detects the temporal patterns of each topic by adding another time variable. Some researchers have also considered the human factor, such as sentiments [64], and also smart-phone locations [28]. These state-of-the-art works have been certified to be successful in both theory and practical applications, such as those where a real-world task of tracking a volume of terms is to be realized [24]. However, their representation of a Web event is based mainly on keyword vectors. Actually, more information is hidden in the webpages [63] of Web events and there is a better way to preserve their semantics [62]. At the same time, they all rely on time series data. If the data available is limited, the prediction no longer works.

The studies on social environment/context deal mainly with Sociology. How to use AI techniques to resolve the social problems is also known as computational social science [72]. During the life time of humans, any move of a human, a group or an organization will inevitably be influenced by a social environment/context [56]. Some research work has already been carried out on modeling the social environment’s influences on different features of people, such as stress [14] and psychiatric disorders, and quantity and quality of voluntary labor contributions [33]. The social environment/context is connected with the development of technology, and an abstract social environment model is built for the social system, in which operations or actions of members are analyzed [66]. Some properties of social environment are analyzed [2], which are very important for modelling social environment. All these researches fall in the Sociology area and suggest the influence of social environment on our lives. There are indeed some works on the analysis of the social context in social science. However, these state-of-the-art works focus mainly on the characteristics of social context, but not on building a computational model. Therefore, they cannot be directly used for prediction of Web event evolution.

To sum up, it is now known that the traditional temporal patterns mining methods or algorithms do not work well if the Web event does not have adequate historical data. It is also known that the social environment influence peoples’ behavior in society. A natural idea that emerges from this knowledge is that we can utilize the social environment to predict the future trends of Web events. To accomplish that, the social environment should be modelled in a computable way to quantify its influence. However, the scope of social environment is too wide a concept to be covered here in its entirety. Therefore, the scope of this paper is restricted to discussing its impact on the evolution of Web events. Since the essence of the social context is social environment, its properties, such as associated network property and hierarchy property, will be considered during the modelling of the social context. Thereafter, we will explain our idea in greater detail.

3 Problem formalization

A Web event is an event that attracts a lot of public attention and there are many webpages published to cover it on the web. Illustrating the evolution of a Web event are the contents of its webpages, which differed on each day in tune with the social context. In order to predict the evolution of Web events under the effect of the social context, we need a computational model for a Web event at first. That is,

Definition 1 (Event Keyword Link Network (EKLN), Ω) An Event Keyword Link Network, Ω, which is composed of the keywords and their associated relations in a Web event, is used to represent the content of a Web event at a specific time and can be formalized as,

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

where K_t^e is the keyword set of a Web event e at time t with weights, R_t^e is the association rule set of a Web event e at time t. An illustrative example of EKLN is shown in Figure 1.

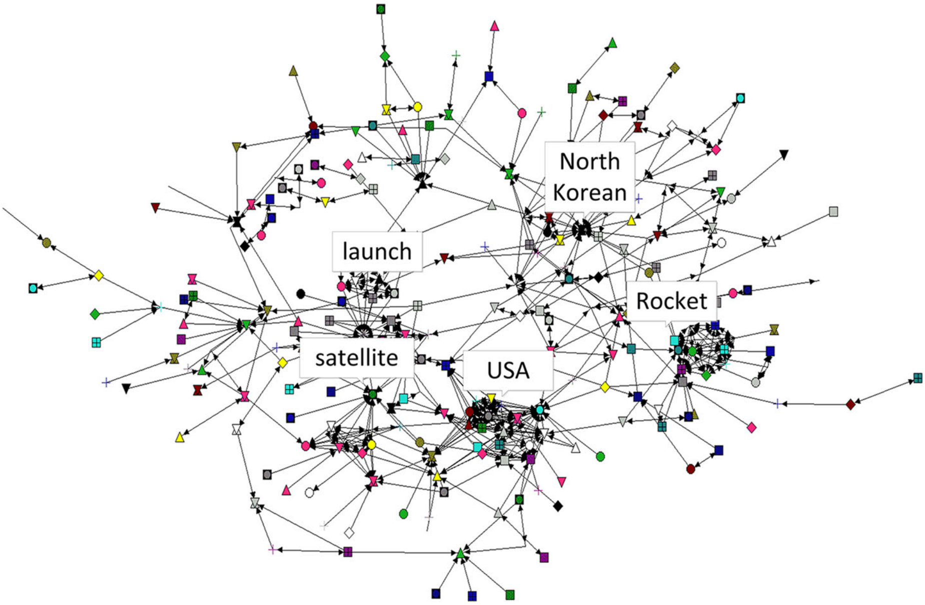

Figure 1 An example of Event Keyword Link Network (Ω) of a Web event, *North Korea launch rocket*, on a given day. This Ω, including its nodes, links and structure, will change/evolve with time under the social context

It should be noted that we use keywords and their association rules (Event Keyword Link Network) as tools to describe the contents/semantics of a Web event (and the social context in next Section) following the commonly accepted assumption in a semantic text mining area [6, 59]. This assumption states that keywords are the basic semantic elements to express the semantics of texts. Here, we apply them to capture the semantics/contents of a Web event and the social context formed by many historical Web events. Except for the keywords, we also consider the association rules between keywords and the network formed by linking keywords with these association rules.

Let us assume that we have a collection of webpages about a Web event e at and only at time t and a volume of past Web events in the human society. Then, this Web event at time t can be represented as Ω_t^e. *The problem is how to predict the change/evolution of a Web event, $\Delta(\Omega_t^e, \Omega_{t+1}^e)$, between Ω_t^e and the unknown Ω_{t+1}^e under a social context with only Ω_t^e in hand?* $\Delta(\Omega_t^e, \Omega_{t+1}^e)$ denotes the change/evolution of Web events, which includes the change of nodes (i.e., emerging or missing nodes), the change of links (i.e., emerging or missing links) and the change of structure (i.e., the increasing or decreasing clustering coefficient). If we have Ω_{t+1}^e and Ω_t^e, $\Delta(\Omega_t^e, \Omega_{t+1}^e)$ can be quantified by different existing methods [49]. It should be noted that, due to the 'burst' property of Web events, it would be assumed that there is no information about this Web event at time Ω_{t+1}^e. This assumption covers two situations: 1) Historical data about Web events is limited when the events break out; and 2) the former temporal pattern cannot help in predicting the future trend because of the 'burst' property of Web events. These two situations provide little information for predicting future trends, which forcing us to seek the help of the social context.

As shown in Figure 2, we propose to resolve this issue as follows: 1) the content/semantics of a current Web event is captured through EKLN in Definition 1; 2) the

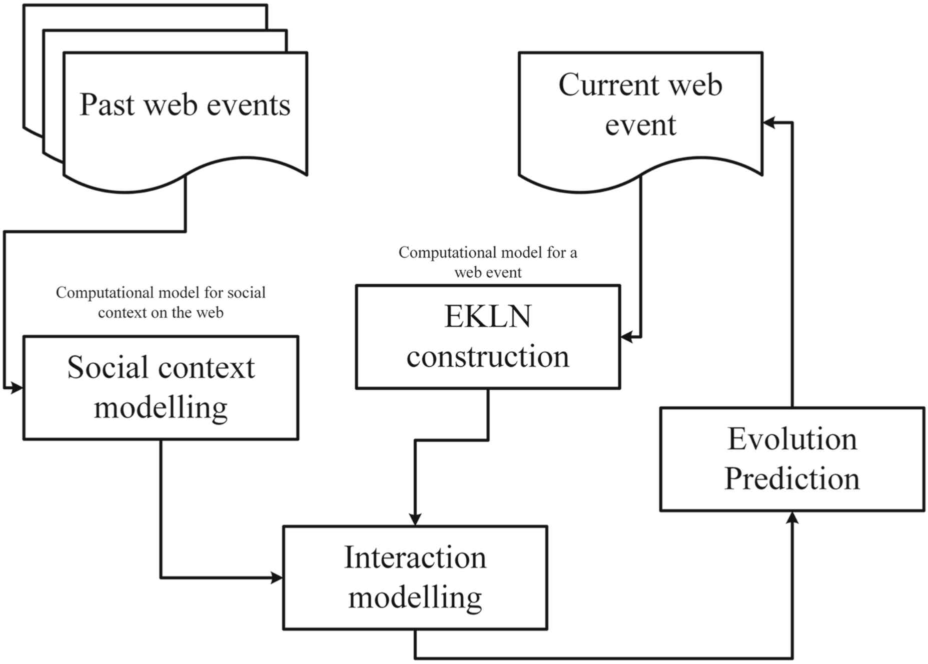

Figure 2 The framework of the proposed method

social context is modeled to make it computable (Section IV); 3) the interaction between social context and Web event is modeled and the impact on the evolution is quantified to make prediction of the near future (Section V). The notations used in this paper are summarized in Table 1.

4 Modelling the social context

This work aims to predict the change of the public interests or attention on a Web event. Our assumption is that the social context will impact on this change, so we need to model social context first. Before starting to build a computational model for the social context, we review the studies in sociology about the social context. In Section IV-A, we list four properties identified in sociology, and these properties have inspired us to propose the network model for the social context in Section IV-B. The social context of a Web event should be the whole web, but it is too difficult and impractical to model the whole web. What significantly influences the evolution of a new Web event are the past Web events. In other words, relative to a new Web event, the atmosphere formed by past events becomes the Web event's social context. We propose a network model to capture the public interests and concerns in history through semantically linking historical Web events. Finally, the temporal dynamic of the proposed social context model is discussed.

4.1 Properties of the social context

Before starting to build a computational model for the social context, we review the studies in sociology about the social context, and find its following properties:

Table 1 Notations of this paper

Notation	Description
e	a Web event
Ω	event keyword link network
Ω_t^e	event keyword link network of event e at time t
K_t^e	keyword set of event e at time t
R_t^e	keyword relation set of event e at time t
$\Delta(\Omega_t^e, \Omega_{t+1}^e)$	predicted change/evolution of event e between time t and $t+1$
$CR(\Omega_t^e, \Omega_{t+1}^e)$	benchmark change/evolution of event e between time t and $t+1$
$\aleph$	social context
ω_k^e	weight of keyword k in event e
$\omega_{i,j}^e$	weight of the association between keyword k_i and k_j in event e
d^e	the network degree of event e in $\aleph$
π^e	the number of webpages of event e
$\mathcal{A}^e$	attractive degree of event e
$s(t)$	Forgetting Curve of Social Event
E^e	the set of activated historical Web events
$\mathcal{C}^e$	the clustering coefficient of event e in the subgraph composed by E^e
$F_{E,t}^e$	external force of event e at time t

1. **Associative Network Property** Because of this property, the constituent elements of the social context are associated with each other. This property derives from the results of the Sociology [17].
2. **Hierarchy Property** Because of this property, each element in the society has its social position, and these social positions, in turn, have their hierarchical structure [37].
3. **Small-World Property** This relates to the renowned ‘Six Degrees of Separation’, which are from the empirical study of Michael Gurevich [22] and of de Sola Pool [58] and some experimental studies of Stanley Milgram [45].
4. **Scale-Free Property** Another common phenomenon of complex networks is being ‘scale-free’ [4]. The ‘associative network’ of social context also has this phenomenon [3, 18].

Note that four computational properties of the social context are listed here for the Web event evolution analysis. However, we do believe that some more properties, such as the family climate [41] and stress [21], do exist, but they are used mainly to investigate personal behavior in the social science, and they do not contribute much to the evolution of Web events. We will consider those properties in future, if we can obtain the social network information [34, 60] and the profile of each person [32] in a Web event. But, for this study, we focus only on the four main properties, mentioned above. Inspired by the above-mentioned properties of the social context in sociology, we have an idea to build a information network as the computational model of social context of Web events in the following section.

4.2 A computational model for the social context

Considering the properties of the social context, Association Linked Network (ALN) [38] is adopted here for building the computational model of the social context. In ALN, different Web events in the society are linked by their associated relations, based on their content, i.e., keywords and keyword associations. If one event is activated by an emerging Web event, the probability of other events, which have a link with this activated event, also being activated is large.

Definition 2 (Social Context of Web Events, $\aleph$) A social context, shown in Figure 3, is an organized structure of human activities (historical Web events) on the web, which is composed of two layer networks: Web event layer network and keyword layer network. The social context can be activated by the current Web event; the former, in turn, can impact the evolution of the latter. It can be formalized, thus:

$$\aleph = \left\langle \begin{array}{c} \{E, R\} \\ \{K^{e_i}, R^{e_i}\} \end{array} \right\rangle \tag{2}$$

where E is the set of historical Web events and R is the set of associations between historical Web events.

Equation (2) is designed to capture the properties of the social context analyzed in Section 4.1. For example, a two-layer-network structure is used to preserve the Hierarchy Property of the social context. Associations between items are used to preserve the Associative Network. ALN construct strategy [38] is used for Small-World and Scale-Free properties of the social context.

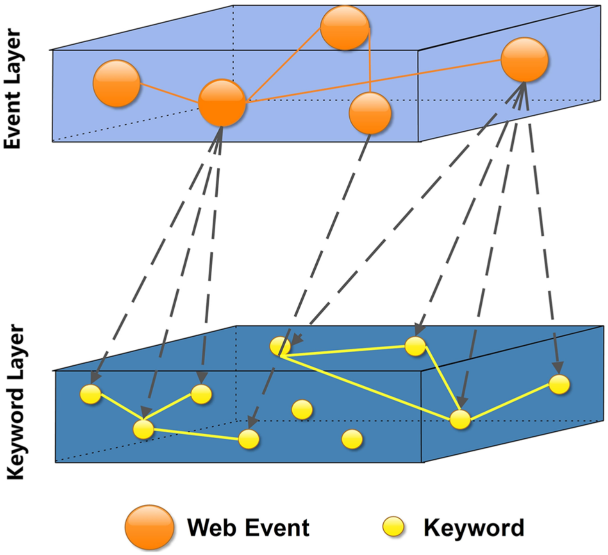

Figure 3 Social context represented by two layer networks, namely event layer and keyword layer

Algorithm 1 Social context construction.

```
Input: A set of historical Web events $\{e\}$ with their corresponding webpages
Output: Social Context, $\aleph$
Extract keywords $K^e$ for all Web events [54]; ;
Extract the association rules $R^e$ between keywords for all Web events [1]; ;
for each Web event e do
    Link the keywords by association rules to form keyword network;;
    for each pair of webpages do
        Compute its association rules as $\sum_{<k_i,k_j>} \omega_{i,j}$, where $< k_i, k_j >$ are two
        keywords from two webpages;
    end
    After normalization, link webpages by their association rules to form webpage
    network;;
end
for each pair of Web events $< e_i, e_j >$ do
    Compute its association rules as $\sum_{k_i \in e_i, k_j \in e_j} \omega_{i,j}$;
end
After normalization, link Web events by their association rules to form Web event
network.
```

A large-scale sample of past Web events and their corresponding webpages are selected to construct $\aleph$. At each layer, the nodes (i.e., Web events or keywords) are connected by their associated relations to simulate the '*association network*' of $\aleph$. The whole construction algorithm is shown as Algorithm 1. Each keyword of a given Web event has a $tfidf$ weight [54],

$$\omega_k^e = tf(k) \times \log\left(\frac{N^e}{N_k^e}\right) \tag{3}$$

where $tf(k)$ is the number of keyword k shown in the webpages of Web event e, N^e is the webpage number of e, and N_k^e is the number of webpages containing keyword k in e. The weight of each association rule between keywords is given by [1],

$$\omega_{i,j}^e = \frac{N_{k_i,k_j}^e}{N_{k_i}^e} \tag{4}$$

where N_{k_i,k_j}^e is the number of webpages containing keywords k_i and k_j.

The time complexity of Algorithm 1 is given below. At the keyword level, the complexity of the keyword extraction for each event is $O(DK)$ where D is the number of documents of an event and K is the number of keywords of this event. The complexity of the keyword relational rules is $O(K^2)$. For all events, the complexity is $O(EDK + EK^2)$ where E is the number of events. At the event level, the complexity for event relationship learning is $O(E^2)$. To total time complexity is $O(EDK + EK^2 + E^2) \approx O(EDK + EK^2)$ because $K >> E$ normally. Note that the construction of computational social context model is one-off and can be computed in an offline fashion.

The keywords of Web events are kept as the 'clues' for association with new Web events. These 'clues'/keywords are very important for the interaction between a current Web event and the social context and the prediction of 'bursty'. For example, the probability of a keyword, with a large clustering coefficient, making a Web event 'burst', will be very high. Actually, these clues connect/associate the Web event with the social context. This aspect will be further discussed in the following Section.

Definition 3 (Web Event in Social Context) Each node in the event layer of the social context is a past Web event with some properties,

$$\left\{t, \mathcal{A}^e, K^e\right\} \tag{5}$$

where t is the time of this event, K^e the keyword set and $\mathcal{A}^e$ the attractive degree of this event e, which is defined as,

$$\mathcal{A}^e = d^e \cdot \pi^e \tag{6}$$

where d^e is the network degree of event e at the event layer ALN of $\aleph$, which expresses the association power to other events, and π^e is the number of webpages covering this event.

Attractive degree, $\mathcal{A}^e$, is a measure of different weights of past Web events in the social context $\aleph$. d^e and π^e reflect $\mathcal{A}^e$ from different perspectives and d^e derives from the network structure view. Due to the '*Associative network*' property of $\aleph$, large value of d^e means that the current Web event may be associated with many past Web events through e, and that the influence of social context on the evolution of the current Web event is large. Since the degree of nodes in networks is the most basic property, d^e is adopted to describe its network structural weight. π^e derives from the volume view; it is also an expression of its power to influence the evolution of a current Web event. If an old event has a large degree of attractiveness, the probability of this Web event having a large evolving influence on current Web events will be high. Over a period of time, there will be certain hot items with high attractive degrees in the social context that attract a lot of attention. If an emerging Web event is associated with one or more of such hot items, the evolution process will be apparently different from the one that is not associated. The weights of different items in $\aleph$ are expressed by the attractive degree $\mathcal{A}^e$, which, in turn, is expressed by two factors: the network degree d and the number of covering webpages π.

To summarize, we use the ALN technique to model social context, based on the analytical outcome of its properties. Although some alternatives are available for semantic representation of social contexts, such as Vector Space Model (VSM) [55] and Ontology [42], they have not been used for this issue. VSM uses a vector to represent an object, and the nature of that vector renders is easily computable. Ontology is a rich semantic representation model with accurate relations between different attributes of an object. The proposed ALN has the following advantages in comparison to VSM and Ontology: 1) Unlike Ontology, ALN can be automatically constructed using Algorithm 1; 2) ALN can express more semantics than what VSM can (the quantitative comparison is given in Section VI-D); and 3) ALN can preserve more properties of the social context of Web events.

Note that the proposed multi-layer network model for the social context is not a multi-layer social network [9, 15] where each node denotes a user and edges are multi-relations between users. Our model is more like a multi-layer information network which is an instance of the general multi-layer network [8, 29], where nodes are Web events or keywords.

4.3 Forgetting mechanisms of the social context

Not only the emerging Web event, but also the social context, evolves with time. This temporal evolution reflects the changing interests of the people in the society. Therefore, we propose a strategy to simulate this evolution to make the social context model more realistic. The idea for the strategy is drawn mainly from Human Memory Update Theory [47]. The social context can be seen as the collective memory of the public [48], so that the human memory decline can be simulated as the decline of social context. Hermann Ebbinghaus proposed, in 1885, the hypothesis of the forgetting curve, which describes the decline of memory retention with time. This curve shows how information is lost over time if there has been no event to recall it [36]. In other words, if memory has not been recalled for a long time, it will be very difficult to recall it later.

The temporal change/evolution of the social context, as illustrated in Figure 4, can be inferred from the change in the weights of items, i.e., past events or keywords. Their weights should be reduced if they have not been mentioned by the public for a long time, because

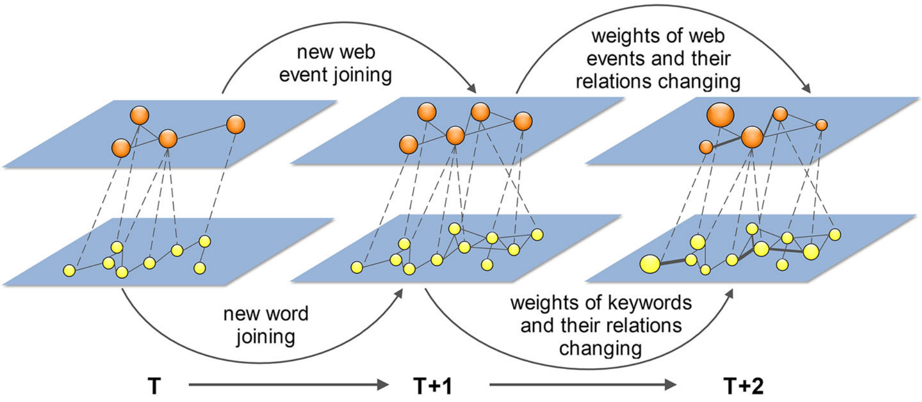

Figure 4 Illustration of social context evolution day by day. Orange nodes denote Web event in social context; yellow nodes denotes keywords; the size of a node denotes its weight; the width of a link denotes its weight

people have lost interest in them (people no longer tend to pay attention to them). Then, those items rarely show any association with the emerging Web events and, consequently their influence will be small, as compared to the influence of other items. To quantify this temporal degeneration, we use the Forgetting Curve of Human Memory [47] (see Figure 5), and to quantify the change of weights of past events in the social context, the Forgetting Curve of Social Event:

$$s(t) = 0.2e^{0.42/(t+0.00255)^{0.225}}. \tag{7}$$

Based on this temporal decline function, we update the attractive degrees of events in the social context, using

$$\mathcal{A}^e(t) = \mathcal{A}^e \cdot s(t) \tag{8}$$

where t is not a specific time point, but a time span between the emergence time of this event and the current time. This paper sets t as the interval (number of months) between two time points. This function has the property of converging to the fixed value 0.2, when t is infinitely great. This implies that an event in the society does not vanish completely, but continues to exist in the society with a very small weight.

If a new Web event activates certain items in the social context, i.e., events and keywords, the weights of those items will be enhanced. This is a 'recall' process and also a memory enhancement procedure. This recall will remind people about these former items. These items can then be easily associated with future Web events, and the influence of these items will be bigger than what it was then they were,

$$\mathcal{A}^e(t) = \mathcal{A}^e \cdot s(0) \tag{9}$$

After introducing the two mechanisms of losing and enhancing, we can simulate the temporal evolution of the social context by adjusting the weights of items.

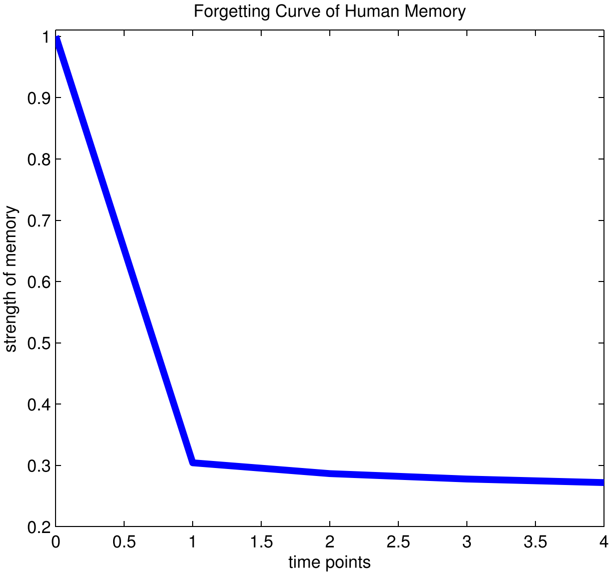

Figure 5 Forgetting curve of human memory

5 Social context-based Web event evolution analysis

Having constructed the social context, we will now discuss how to quantify the influence of social context on the evolution of Web events. First, we model the interaction procedure between a Web event and its social context, which gives some hints on the quantified influence of the social context. Then, we introduce the equation to quantify in the influence of the social context in detail.

5.1 Interaction between the social context and a Web event

To model this interaction, we resort to the Anchoring Theory in Psychology [12, 19]. Anchoring is a pervasive judgment bias in which decision makers are systematically influenced by the starting point [12, 20, 46]. For example, let it be assumed that the people are first informed of a random number, say *400*. Then, they are asked to hazard a guess about something, say, the height of a building. The results will tend to be close to *400*, even though those people know that there is no relation between *400* and the height of the building. This phenomenon is called mental anchoring; it reflects the influence of the anchor on the judgement of decision makers. In anchoring theory, two factors impact the degree of influence: 1) the number of anchors that cause activation; 2) the strengths/weights of anchors.

We believe that the keywords play the role of 'anchors' during the interaction between the social context and a Web event. The whole interaction is modelled as follows:

1. **Activation**-Certain keywords in the keyword layer network of the social context are associated with the keywords of a new Web event; as a result, the corresponding past events, described by these keywords in the event layer network of the social context are also associated.
2. **Feedback**-After the commencement of 'activation', some items (i.e., past Web events) in $\aleph$ are activated, which would then impact the evolution of a current Web event.

For quantification of influence, the two procedures of interaction, i.e., activation and feedback, can be formalized together, thus:

$$\begin{aligned} E^e &= \left(w_1^e \; w_2^e \; \cdots \; w_K^e \right) \begin{bmatrix} w_1^{e_1} & \cdots & w_1^{e_m} \\ \vdots & \ddots & \vdots \\ w_K^{e_1} & \cdots & w_K^{e_m} \end{bmatrix} \\ &= \left(\sum_k w_k^e \cdot w_k^{e_1} \; \sum_k w_k^e \cdot w_k^{e_2} \; \cdots \sum_k w_k^e \cdot w_k^{e_m} \right) \end{aligned} \tag{10}$$

where w_k^e the weight of keyword k in a current Web event e; E^e is the feedback of past Web events with weights $\{\sum_k w_k^e \cdot w_k^{e_m}\}$.

The left part of (10) represents the 'activation' procedure and the right part the 'feedback' procedure. Equation (10) thus represents the degree of association between a past event and a current event, $\sum_i w_i^e \cdot w_i^{e_m}$, corresponding to the two factors of anchoring: the summation, $\sum_i$, for the number of keywords/'anchors' and the product, $w_i^e \cdot w_i^{e_m}$, for strengths of keywords/'anchors'. The physical meanings of w_i^e and $w_i^{e_m}$ are explained with an example. Let a Web event *North Korea rocket launch* be assumed with only two keywords: *missile* and *research*. These keywords may simultaneously activate two past Web events in the social context: *North Korea nuclear weapon* and *The Nobel Prize in Science*. Different w_i^e, is reflected by the different sizes of *missile* and *research*. It means that the influence of *North Korea nuclear weapon* on the evolution of *North Korea rocket launch* is greater than the influence of *The Nobel Prize in Science*; Different $w_i^{e_m}$, is reflected by the different sizes of *missile* in e_1 and *research* in e_2. It means that the influence of *North Korea nuclear weapon* to the evolution of *North Korea rocket launch* is greater than the influence of *The Nobel Prize in Science*.

Based on the weights of events in E^e, we can know which events are associated with the public, when they are exposed to the current Web event.

5.2 Computation of the external force

For a Web event at time t, certain keywords of this Web event activate various past Web events in the social context $\aleph$, after which the feedback will impact the state of this Web event at time $t+1$. Although the feedback from (10) represents just a set of past Web events, relations do exist between these Web events in the social context. Thus, the feedback is actually a sub-graph of the event layer of $\aleph$, which is just the source of the external force. Accordingly, the definition of external force is given as follows:

Definition 4 (External Force, F_E) External Force is the influence of the social context $\aleph$ on the evolution of a Web event, which can be evaluated by quantifying the feedback sub-graph composed by E_t^e that is the activated historical events in $\aleph$ by current event e at time t.

F_E can also be regarded as the cumulative force of the past Web events and the current Web event. To compute external force, some factors that impact the quantification of this subgraph are described first:

1. *The attractive degrees of activated events in the social context* The attractive degrees that represent the weights of past Web events in the social context, change/evolve with time and affect the evolution of the current Web event are discussed under Section 4.3;
2. *The number of activated past Web events in the social context* The number of activated events in $\aleph$ relates to the extent the current Web event can activate the social context $\aleph$;
3. *The average cluster coefficient of the subgraph* Cluster coefficient [65] of an event, which comes from the association network property of $\aleph$, is discussed under Section 4.1. It refers to the ability of this event in the subgraph to gather its neighbours together. The term ‘together’ means that each activated event has close relationship with its neighbouring events. Apparently, the closer the subgraph, the greater would be its influence on the evolution of the current Web event from $\aleph$;
4. *The weights of keywords that cause association* The weights of these keywords have two parts: one is the weights in the current Web event and the other is the weights in the past Web events in the social context, as discussed in Section 5.1.

As the four foregoing factors are proportional to the value of external force, the product of their values (after proper normalization) is used to compute the value of external force as follows:

$$F_E^e = \sum_{e_i \in N^e} \left(\mathcal{A}^{e_i} \cdot \mathcal{C}^{e_i} \cdot \sum_{k \in K^{e_i}} \sqrt{w_k^e \cdot w_k^{e_i}} \right) \tag{11}$$

where $\mathcal{A}^{e_i}$ is the attractive degree of event e_i, and $\mathcal{C}^{e_i}$ is the cluster coefficient of e_i in the subgraph of E^e, K^{e_i}, which is a set of keywords that cause the activation at time t; ω_k^e is the weight of a keyword of a current Web event e and $\omega_k^{e_i}$ is the weight of this keyword in e_i. The form of $\sum_{k \in K^{e_i}} \sqrt{w_k^e \cdot w_k^{e_i}}$ is due to the symmetry of keywords both in the current Web event and in the social context from (10).

Until now, we could evaluate the influence from the social context to the evolution of the Web events as the external force by (11). The main computational complexity of the proposed method is to activate past Web events by (10) and evaluate the (11). The complexity of (10) is $O(K^2)$. As for (11) where $\mathcal{A}^{e_i}$ and $\mathcal{C}^{e_i}$ can be pre-evaluated and stored so there is no computational burden for them. To evaluate the $\sum_{k \in K^{e_i}} \sqrt{w_k^e \cdot w_k^{e_i}}$, the complexity is $O(K^2)$. The total time complexity is around $O(K^2)$. Next, we will design a series of experiments to validate the effectiveness of this evaluation.

6 Evaluation and discussion

To verify the rationality of our explanation for the evolution of a Web event under the social context, an evaluation metric is developed and tested on the data sets.

6.1 Evaluation metric

As the proposed external force is used to predict the change/evolution between the current state and the next-time state of Web events, the predicted change and the change between two consecutive time stamps of Web events are compared to verify our research results. If the value of external force at t has a large correlation with the change of a Web event from t to $t+1$, the rationality of our work can be verified.

As the Web event is represented as Ω, the change between two consecutive time points, $\Delta(\Omega_t^e, \Omega_{t+1}^e)$, can be evaluated by the similarity of two corresponding complex networks, called Change Reference, CR, for the evolution of a Web event under external force. So, we give a new definition as follows:

Definition 5 (Change Reference, CR) Change Reference is the change/evolution of a Web event e, evaluated from time point t to time point $t+1$, based on the data (two keyword networks, Ω_t^e and Ω_{t+1}^e) at two time points.

The computation of CR is based on the following Vector Similarity Algorithm [49], which jointly considers the nodes and edges of keyword networks.

A graph (EKLN here) is represented by two vectors: the node vector and the edge vector. The similarity between two graphs (EKLN) is computed as,

$$CR = \frac{CR^{node} + CR^{edge}}{2} \tag{12}$$

where CR^{node} and CR^{edge} are the similarities between node and edge vectors of two graphs, respectively. They are computed as,

$$CR^{node} = \frac{\sum_{(n,m) \in V^t \cup V^{t+1}} \frac{|w_n - w_m|}{\max(w_n, w_m)}}{|V^t \cup V^{t+1}|} \tag{13}$$

and

$$CR^{edge} = \frac{\sum_{(u,v) \in E^t \cup E^{t+1}} \frac{|w_u - w_v|}{\max(w_u, w_v)}}{|E^t \cup E^{t+1}|} \tag{14}$$

where V^t and V^{t+1} are node sets of EKLNs at two time points and w_n is the weight of node n; E^t and E^{t+1} are edge sets of EKLNs at two time points and w_u is the weight of edge u.

The changes for every two consecutive Ω's of a number of Web events are computed and the counterpart external forces by (11). The Pearson Correlation Coefficient [5] is adopted here to quantify the correlation between the values of $CR : \{CR_1, CR_2, \ldots, CR_{n-1}\}$ and $F_E : \{F_{E,1}, F_{E,2}, \ldots, F_{E,n-1}\}$, where n is the day number of a Web event, CR_t is the CR between day t and day $t + 1$, and F_t is the F_E at day t.

Actually, external force is computed to predict the value of CR, which is the final aim of this work. For example, $F_{E,1}$ is computed based only on the data at time t_1, to predict the change/evolution CR_0 between two times t_1 and t_2. In the designed experiment setting, with data at two time points t and $t + 1$ in hand, we first compute the external force $F_{E,t}$, based on the data at time point t alone, and then compute CR_t, based on the data at two time points t and $t + 1$ together.

Since predicting the exact value of CR is too hard, we do not expect that $CR = F_E$, but we do expect a high correlation between them. And, if so, it means that our algorithm of external force computation is reasonable and can be used to predict the change/evolution of Web events. This correlation suggests that we can predict the next-step, namely the change/evolution of a Web event, based on the analysis of the data at a current time point alone.

It should be noted that, although the whole life cycles of Web events are used as the data, the computation of external force at any given time stamp does not use the data at other time stamps, but only the data at that current time stamp. This means that the forces at different time stamps are independent of each other. Therefore, it is different from the time series-based approaches.

6.2 Datasets

Our research has used two datasets, both composed of Web events, to validate our proposed approach. One dataset is used to construct the social context and the other is regarded as emerging Web events that interact with the social context. Both datasets are Chinese webpages; of course, our approach is not language-dependent (the only difference between different languages is in the Natural Language Processing software chosen to extract words of a specific language). Chinese webpages are chosen because China has many Web events by virtue of its population size and speed of economic growth.

6.2.1 Dataset I

The Web events used for constructing the social context $\aleph$ are selected from everyday hot Web events provided on the *Baidu News*,[3] and downloaded as past Web events.

It may be noted that the selected Web events have been hot events in China and there are already some relevant webpages (normally 2-5 days) on those events. So, we identify the very starting date of each Web event. There are some works that deal with determining the starting date of a Web event, such as [26]. To ensure accuracy in identification, we first read all the collected webpages about a Web event and select a webpage with the condition of having small date and big relevance. While doing so, we know the exact date on which a Web event occurs in the real world, such as the date of the *Japan Nuclear Leakage*. Considering these two factors, we finally determine the starting date of a Web event. The ending date is determined by the collected webpages. The last date is set as the ending date.

[3] http://news.baidu.com/

Table 2 Dataset I

Names of variables	Value
Number of Web events	20575
Number of Webpages	17,595,529
Number of Keywords	1,555,364

The related webpages about a Web event are downloaded through Web search engines, such as *Baidu* (www.baidu.com), the largest search engine in China and *Google* (www.google.com.hk). The collection procedure is as follows: 1) Select a Web event with a name, like *Japan Nuclear Leakage*; 2) Feed the name of the Web event and the date into the search engines; 3) Obtain the results returned from search engine (usually with multiple pages); 4) Download top 500 related webpages[4]; 5) Feed a new date into the search engine and do Step 2.

The time range of these Web events is June 29, 2009 to March 27, 2012. These events are of utmost concern to the Chinese society in their day-to-day life. Each event is represented by 50 keywords and their associated relations mined from its corresponding webpages. Only that data in the system, which pertains to the required time ranges, is used in our experiments. Note that the process of webpages includes: part-of-speech and lemmatization using the Stanford CoreNLP.[5] Some statistics relating to filtering are shown in Table 2.

6.2.2 Dataset II

The events used as current Web events are collected as Dataset I. The webpages of each Web event are downloaded and analysed, including keyword extraction and associated relations mining. Because our social context relates to the Web events in China, Chinese society forms our research background. The Web pages collected for each Web event start on the initial day identified by human selection. The collection procedure is the same as the one followed for the webpages in Dataset I. Statistics are shown in Table 3. It is to be noted that there is no overlap between the two Datasets, and the starting times of the Web events in Dataset II are later than those in Dataset I. In other words, the Web events of Dataset II follow those of Dataset I. This ensures that the Web events of Dataset I can be considered as the social context for the Web events of Dataset II.

6.3 Experimental results analysis

In this Section, we evaluate the efficacy of our approach in predicting the future trend of Web events by comparing it with a time series-based prediction approach (i.e., ARMA [11]; its implementation in Matlab is adopted).

We compare the performances of F_E with ARMA in terms of predicting the evolution of Web events. Two examples (two Web events) are shown in Figure 6. For each Web event, we compute the F_E and ARMA values and evaluate their correlation coefficients with those of the CR. The x-axis denotes time points. The values of F_E and ARMA at time t are computed based on the data at time $t-1$ and at time $t-3$, $t-2$ and $t-1$, respectively. It is

[4] Note that not all of the webpages can be downloaded because some of them may be videos and some of them cannot be linked and crawled.

[5] http://nlp.stanford.edu/software/corenlp.shtml

Table 3 Dataset II

Names of variables	Value
Number of Webpages	609,696
Number of date points	3,880
Number of Web Events	99
Average day number of a Web Event	40

to be noted that the comparison starts at the third time point because ARMA cannot predict the starting (here, first and second) points. It can be seen that the bigger the correlation with CR, the better is the performance of our approach. From Figure 6, it can be seen that the correlation between F_E with CR is bigger than the values from ARMA with CR. It should be noted that ARMA is based on the data of three former time points, whereas our approach is based only on the data of one former time point. Even under this ‘unfair’ condition, our approach performs better than ARMA does.

The results of all the Web events of Dataset II are shown in Table 4 and Figure 7. To show their statistical significance, the corresponding p-values of the correlation coefficients are also shown in both the Table and the Figure. It can be seen that, comparing with the ARMA, F_E has relatively bigger correlation coefficients with CR on most of the Web events. To compare ARMA and F_E, we mark all the events with big correlation coefficients (≥ 0.5) and p-values (< 0.05). The percentage of marked events is significantly higher than the percentage of ARMA as shown (see Table 4).

An interesting point in Table 4 is that the small correlation coefficients from the proposed method tend to have big p-values (bigger than 0.05). If the correlation coefficients are with

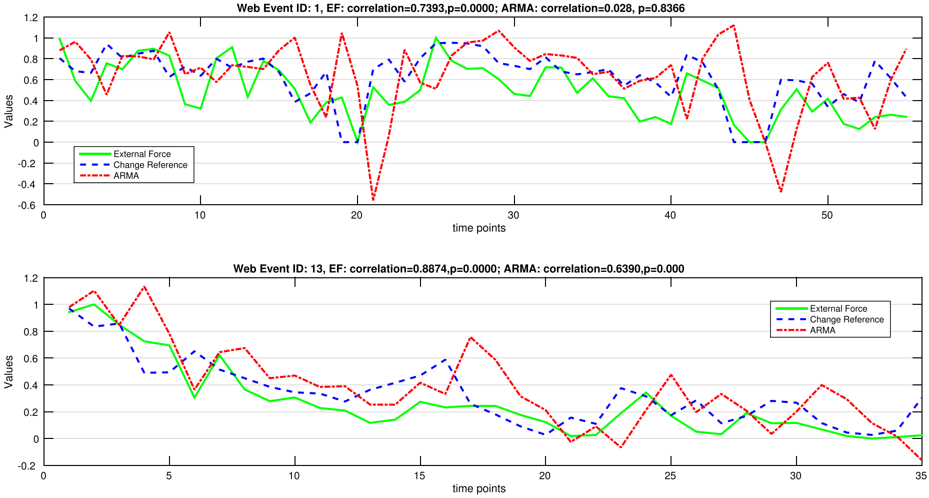

Figure 6 Correlation Coefficients of two Web events between CR with F_E and between ARMA with CR. The x-axis denotes the time. The continuous (green) line represents the values of F_E, the dashed (red) line the values of ARMA and the dotted (blue) line the values of CR. The two Web events are *Libya war* with ID 1 and *Fukushima power plant explosion* with ID 13. It is to be noted that the value of F_E, at a given time t, is based only on the data at time $t-1$, although the whole life of this Web event is used as the dataset. However, ARMA forecasts the value at time t, based on the data at time $t-1$, $t-2$ and $t-3$. Thus, there is a shift of curves in the third and fourth subfigures

Table 4 The correlation coefficients and p-values of Web events from two methods

	External force		ARMA			External force		ARMA	
ID	COCO	P-value	COCO	P-value	ID	COCO	P-value	COCO	P-value
1	0.7393	0.0000	0.0285	0.8366	51	0.5601	0.0002	0.2217	0.1938
2	0.4320	0.0054	−0.3480	0.0348	52	0.6059	0.0000	−0.2326	0.1722
3	0.7769	0.0000	0.1679	0.3584	53	0.6726	0.0000	0.0043	0.9802
4	0.7049	0.0000	0.0920	0.6416	54	0.7326	0.0000	0.3846	0.0171
5	0.9342	0.0000	0.7205	0.0007	55	0.5245	0.0004	−0.2110	0.2035
6	0.6302	0.0001	−0.3173	0.0999	56	0.5437	0.0005	−0.2359	0.1792
7	0.6882	0.0000	−0.0566	0.7707	57	0.4559	0.0040	−0.4012	0.0169
8	0.4594	0.1552	−0.4571	0.2548	58	0.3637	0.0165	−0.0188	0.9085
9	0.2839	0.1690	−0.3203	0.1462	59	0.2981	0.0256	−0.1600	0.2526
10	0.6878	0.0001	0.2641	0.2124	60	0.2692	0.1124	0.2204	0.2177
11	0.5247	0.0000	−0.2202	0.1131	61	0.4453	0.0007	−0.0599	0.6764
12	0.9139	0.0000	0.5767	0.0497	62	0.4804	0.0054	−0.4322	0.0192
13	0.8874	0.0000	0.6390	0.0000	63	0.7155	0.0000	−0.0150	0.9398
14	0.6811	0.0000	−0.0874	0.5919	64	−0.8026	0.0092	−0.4833	0.3315
15	0.6724	0.0000	−0.1371	0.3425	65	0.2842	0.1428	−0.2742	0.1846
16	0.0321	0.8763	0.0351	0.8738	66	0.4848	0.0140	−0.3591	0.1008
17	0.6424	0.0005	0.4745	0.0257	67	0.4514	0.0235	−0.3003	0.1745
18	0.6478	0.0000	−0.0296	0.8541	68	0.2691	0.1125	−0.1573	0.3821
19	0.5406	0.0003	−0.1550	0.3528	69	0.4151	0.0086	−0.4252	0.0097
20	0.6973	0.0002	−0.3445	0.1261	70	0.3593	0.0267	−0.1739	0.3177
21	0.1398	0.5455	0.1006	0.6912	71	0.4421	0.0113	−0.4942	0.0064
22	0.7291	0.0000	0.1425	0.2905	72	0.6970	0.0000	−0.1037	0.5531
23	0.7690	0.0000	0.3495	0.0040	73	0.4642	0.0026	−0.0887	0.6018
24	0.5914	0.0000	−0.2034	0.0988	74	0.3317	0.0047	−0.1419	0.2482
25	0.6430	0.0000	0.2026	0.1028	75	0.6997	0.0000	0.2865	0.0179
26	0.4213	0.0010	−0.3547	0.0079	76	0.4223	0.0144	−0.2538	0.1759
27	0.5692	0.0000	−0.2021	0.1248	77	0.4851	0.0031	−0.2273	0.2108
28	0.5471	0.0000	0.1683	0.1768	78	0.1944	0.3033	−0.4796	0.0114
29	0.1974	0.1411	−0.2683	0.0498	79	0.5149	0.0026	−0.2995	0.1145
30	0.4822	0.1581	−0.0380	0.9355	80	0.5367	0.0013	−0.3921	0.0321
31	0.0153	0.9285	0.1509	0.3944	81	0.5461	0.0022	−0.3072	0.1269
32	0.1885	0.1103	0.1166	0.3364	82	0.5526	0.0019	−0.3976	0.0443
33	0.0056	0.9693	−0.4495	0.0017	83	0.4275	0.0207	−0.5536	0.0033
34	0.6410	0.0001	−0.5626	0.0015	84	0.2263	0.2293	−0.3838	0.0481
35	0.6128	0.0000	0.1915	0.1207	85	0.3678	0.0496	−0.3727	0.0608
36	0.6773	0.0000	0.4063	0.0014	86	0.5506	0.0016	−0.4556	0.0169
37	0.6713	0.0000	−0.0107	0.9394	87	0.5302	0.0022	−0.3919	0.0392
38	0.3078	0.1999	0.2708	0.3103	88	0.2016	0.2605	−0.4564	0.0112
39	0.4723	0.0000	−0.2190	0.0706	89	0.5177	0.0014	−0.2703	0.1347

Table 4 (continued)

	External force		ARMA			External force		ARMA	
ID	COCO	P-value	COCO	P-value	ID	COCO	P-value	COCO	P-value
40	0.7103	0.0000	0.1705	0.1779	90	0.3290	0.0616	−0.3792	0.0387
41	0.6264	0.0000	−0.2364	0.1318	91	0.5373	0.0011	−0.2211	0.2320
42	0.7501	0.0000	0.5094	0.0048	92	0.5645	0.0005	−0.2802	0.1269
43	0.6102	0.0093	−0.3144	0.2737	93	0.1636	0.3478	−0.2229	0.2202
44	0.5354	0.0001	−0.0089	0.9545	94	0.6837	0.0000	−0.1408	0.4345
45	0.6143	0.0000	0.1963	0.2443	95	0.0731	0.7169	−0.2917	0.1666
46	0.6116	0.0000	0.4736	0.0020	96	0.6588	0.0002	−0.4402	0.0313
47	0.3160	0.0065	0.0306	0.8015	97	0.5471	0.0031	0.4337	0.0342
48	0.7130	0.0000	0.1820	0.2810	98	0.4940	0.0103	−0.3080	0.1528
49	0.3881	0.0122	−0.1822	0.2735	99	0.6358	0.0005	0.1585	0.4702
50	0.6821	0.0000	0.2139	0.2037					

The marked events relate to those whose correlation coefficients are greater than 0.5, and p-value smaller than 0.05

big p-values, then they are untrusted, and hence no conclusion can be drawn, based on them. In other words, it can be concluded that the F_E of most Web events (with p-values smaller than 0.05 and correlation coefficients bigger than 0.5) are statistically correlated with CR, and that the relations between F_E and CR of a limited number of Web events (with p-values bigger than 0.05) are unsure but not statistically uncorrelated.

The reason why most of the Web events in Dataset II have relatively big correlation coefficients, shown in Table 4, is due to our data collection. Since our Web events originate from the recommended topics of *Baidu*, most of them have the potential to attract more attention and to activate past events in the social context, and thus the whole value is relatively large.

The computation of CR is about $O(K^2 + R^2)$ where K is the number of keywords in current Web event, and R is the number of relations in current Web event. Since ARMA needs a time series of CRs of a Web event, the time complexity of ARMA is approximately

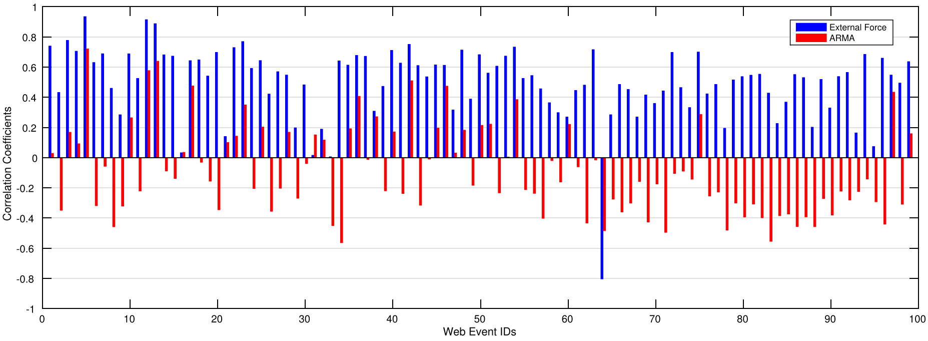

Figure 7 Histogram of the correlation coefficients of all the Web events obtained from the two methods: External Force and ARMA. The corresponding values of correlation coefficients and p-values are given in Table 4

$O(n(K^2 + R^2))$ where n is the number of time points fed into the model. The proposed method contains two parts: one is the activation historical Web events $O(NK)$ where N and K are numbers of historical Web events and keywords in social context; and the other is the computation of (11) in manuscript $O(N^e K)$ where N^e is the number of activated historical Web events. Since N^e is greatly smaller than N, the time complexity of the proposed method is approximately $O(NK)$. The scale of relations (i.e., R) is usually about 10 times of the scale of keywords in event keyword link network. It appears that the proposed method is with lower time complexity when $100nK > N$, and this condition can easily be satisfied. Note that although the proposed method needs the social context model in hand, this model could be built in an offline fashion. Here, an advice is that if they have abundant historical Web events for modelling social context, the time cost of proposed algorithm is smaller than the time-series-based methods.

6.4 Evaluations on model components

To demonstrate the efficiency of the proposed model, we have compared it with the efficiency of some candidate components. The proposed model has broadly two components: (i) social context modelling and (ii) interaction modelling between social context and a Web event. The candidate components for comparison are chosen keeping this composition in view.

For comparing social context modelling, the following are the candidates chosen:

1. **SC-var1** models social context without considering its temporal evaluation. This method will influence the computation of the External Force in (11) through the value of $\mathcal{A}^{e_i}$;
2. **SC-var2** uses the Vector Space Model (VSM) [55] to model the social context. With VSM, there will be no relations between the past events in the social context. Then, the computation of the External Force in (11) will not involve $\mathcal{C}^{e_i}$, which is from the relations between the past events.

These two candidates are used to replace the social context modelling component of the proposed model (the other components are kept unchanged). The results are shown in Figure 8. The first row shows the results from the proposed model and ARMA. The pie charts show the percentages of different events: the blue area indicates the number of events with correlation bigger than 0.5 and p-values smaller than 0.05; the green area indicates the number of events with the correlation within [0.3, 0.5] and p-values smaller than 0.05; the yellow area indicates the the remaining events. The pie chats can be seen as a summary of the Table 4. The second row shows the results from the proposed model with the candidates as components. These results conclusively show that the proposed social context model is better than VSM-based model. Furthermore, temporal evolution is important for social context modelling as we have discussed in Section 4.3.

For interaction modelling between social context and a Web event, the candidates for the comparison are,

1. **IT-var1** A Web event activates the past events considering only the *name distance*, which can be evaluated as the number of matching keywords between the names of current Web event and past events.
2. **IT-var2** A Web event activates the past events in the social context considering only the *time distance*, which can be evaluated as the difference between the current time of the Web event and the emerging time of the past events.

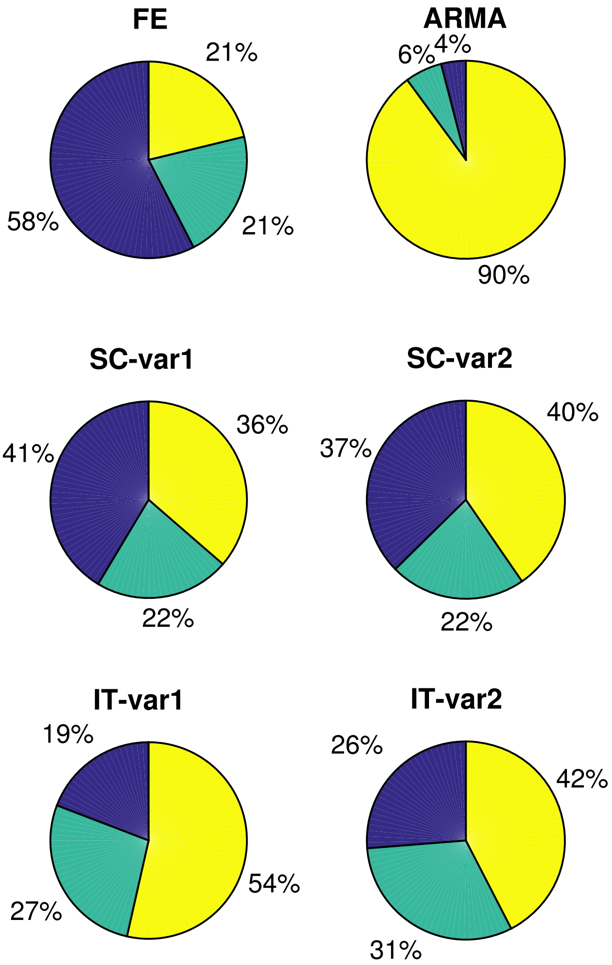

Figure 8 Demonstration of the performances of the components in the proposed model. The pie chart shows the percentages of different events. The total number is 99. The blue area denotes the percentage of the events with correlation bigger than 0.5 and corresponding p-values smaller than 0.05; the green area denotes the percentage of the events with correlation within [0.3, 0.5] and corresponding p-values smaller than 0.05; the yellow area denotes the remaining events

Both these candidates will influence the computation of the External Force in (11) with the part $\sum_{k\in K^{e_i}}\sqrt{w_k^e \cdot w_k^{e_i}}$ replaced by the new distances. As before, the two candidates are used to replace the interaction modelling component of the proposed model (the other components are kept unchanged). The results are shown in the third row of Figure 8. The comparisons show that the original keyword matching strategy gives the best performance, in comparison to the performances of **IT-var1** and **IT-var2**. This is because the method in the proposed model matches the whole content/semantics of the events, rather than merely the names or the times. The influence of the past events with similar content will be larger than the one with similar name or closing time.

6.5 Noise analysis

Initially, we are of the belief that there will be few noise webpages among those downloaded during data collection. Therefore, we do not claim that all the webpages are 100 percent relevant to a Web event. However, considering our data collection strategy, we claim that most of the webpages downloaded are relevant to a Web event. Since the proportion of noise webpages is small, the keywords mined from these noise webpages tend to be weighted with small values by (3) and posited along the edge of the keyword network. Some noise keywords will be posited along the edge of the keyword network of a Web event, because very few webpages support the links obtained by (4) (increase the weights of links) between noise keywords and other main keywords of a Web event. In addition, the weights of these keywords are also small because the webpages obtained by (3) are not much supported. If they are in the social context, they are very hard to be associated with a new Web event; if they are in a new Web event, they do not form the main part of the whole semantics of this Web event. Therefore, these noise keywords will not influence the final results very much.

7 Conclusions and further study

The burst property and limited historical data prevent the traditional time-series-based approaches from predicting the content change/evolution of Web events accurately. This research has proposed a social context-based method, with no need for temporal patterns. To model a social context reasonably well, its properties have been analysed and then reflected by the constructed social context model. Accordingly, the interaction between social context and Web events has been simulated, including activation and feedback procedures, to lay the foundation for quantifying the influence of the social context. Based on that, the external force has been defined and quantified as the influence of the social context on the evolution of Web events. Experiments with real-world Web events data, involving comparison of different strategies, demonstrate the reasonability of the constructed social context model. We want to stress that the constructed computational social context can be used not only for predicting the content change, but also for some other quantitative social science works involving social environment. For the task of content change prediction, we have compared the proposed method with time series-based approach, using the real-world Web event dataset. The comparison reveals that the results achieved by the proposed method are better than those achieved by time series-based approach. In the future, we will consider to jointly model the spreading/propagation and evolution of a Web event because we believe there must be some interactions between two behaviours of a Web event.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under grant nos. 91746203, and by the Australian Research Council (ARC) under discovery grant DP190101733.

References

1. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20Th Int. Conf. Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)
2. Augoustinos, M., Innes, J.M.: Towards an integration of social representations and social schema theory. British Journal of Social Psychology **29**(3), 213–231 (1990)
3. Barabási, A.L.: Scale-free networks: a decade and beyond. Science **325**(5939), 412–413 (2009)

4. Barabási, A.L.: Bursts: The hidden patterns behind everything we do. PLUME (2011)
5. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing, pp. 1–4. Springer (2009)
6. Berendt, B., Hotho, A., Stumme, G.: Towards Semantic Web Mining. In: Horrocks, I., Hendler, J. (eds.) The Semantic Web ? ISWC 2002, Lecture Notes in Computer Science, vol. 2342, pp. 264–278. Springer, Berlin (2002)
7. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning (ICML), ICML '06, pp. 113–120. ACM, New York (2006)
8. Boccaletti, S., Bianconi, G., Criado, R., del Genio, C.I., Gómez-gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., Zanin, M.: The structure and dynamics of multilayer networks. CoRR arXiv:abs/1407.0742 (2014)
9. Bródka, P., Kazienko, P.: Multi-layered social networks. CoRR arXiv:abs/1212.2425 (2012)
10. Cai, H., Huang, Z., Srivastava, D., Zhang, Q.: Indexing evolving events from tweet streams. IEEE Trans. Knowl. Data Eng. **27**(11), 3001–3015 (2015)
11. Chan, N.H.: Autoregressive moving average models. Time Series: Applications to Finance with R and S-Plus, Second Edition pp 23–37 (2010)
12. Chapman, G.B., Johnson, E.J.: Anchoring, activation, and the construction of values. Organ. Behav. Hum. Decis. Process. **79**(2), 115–153 (1999)
13. Cho, H., Varian, H.: Predicting the Present with Google Trends. Tech. rep., Google Inc (2009)
14. Creel, S., Dantzer, B., Goymann, W., Rubenstein, D.R.: The ecology of stress: Effects of the social environment. Functional Ecology (2012)
15. Dickison, M.E., Magnani, M., Rossi, L.: Multilayer social networks. Cambridge University Press (2016)
16. Do, Q.X., Chan, Y.S., Roth, D.: Minimally supervised event causality identification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 294–303. Association for Computational Linguistics, Stroudsburg (2011)
17. Edwards, J.: Retrieving the big society. Wiley Blackwell (2012)
18. Forgas, J.P.: Handbook of affect and social cognition. Psychology Press (2012)
19. Furnham, A., Boo, H.C.: A literature review of the anchoring effect. J. Socio-Econ. **40**(1), 35–42 (2011)
20. Galotti, K.M.: Cognitive psychology in and out of the laboratory. SAGE Publications Inc (2013)
21. Gao, C., Liu, J.: Network-based modeling for characterizing human collective behaviors during extreme events. IEEE Transactions on Systems, Man, and Cybernetics: Systems **47**(1), 171–183 (2017)
22. Gurevitch, M.: The social structure of acquaintanceship networks. Massachusetts Institute of Technology (1961)
23. Hennig-Thurau, T., Walsh, G., Walsh, G.: Electronic word-of-mouth: motives for and consequences of reading customer articulations on the internet. Int. J. Electron. Commer. **8**(2), 51–74 (2003)
24. Hong, L., Yin, D., Guo, J., Davison, B.D.: 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), KDD '11, pp. 484–492. ACM, New York (2011)
25. Jiang, M., Fang, Y., Xie, H., Chong, J., Meng, M.: User click prediction for personalized job recommendation. World Wide Web **22**(1), 325–345 (2019)
26. Jin, X., Spangler, S., Ma, R., Han, J.: Topic initiator detection on the world wide web. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, pp. 481–490. ACM, New York (2010)
27. 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 Conference on World Wide Web (WWW), WWW '11, pp. 257–266. ACM, New York (2011)
28. Kelly, D., Smyth, B., Caulfield, B.: Uncovering measurements of social and demographic behavior from smartphone location data. IEEE Transactions on Human-Machine Systems **43**(2), 188–198 (2013)
29. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. Journal of Complex Networks **2**(3), 203–271 (2014)
30. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. In: Proceedings of the 12th International Conference on World Wide Web (WWW), pp. 568–576. ACM, New York (2003)
31. Lee, P., Lakshmanan, L.V., Milios, E.: Keysee: Supporting keyword search on evolving events in social streams. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, pp. 1478–1481. ACM, New York (2013)
32. Liang, G., He, W., Xu, C., Chen, L., Zeng, J.: Rumor identification in microblogging systems based on users' behavior. IEEE Transactions on Computational Social Systems **2**(3), 99–108 (2015)
33. Linardi, S., McConnell, M.A.: No excuses for good behavior: Volunteering and the social environment. J. Public Econ. **95**(5), 445–454 (2011)

34. Liu, G., Liu, Y., Liu, A., Li, Z., Zheng, K., Wang, Y., Zhou, X.: Context-aware trust network extraction in large-scale trust-oriented social networks. World Wide Web **21**(3), 713–738 (2018)
35. Liu, Y., Xu, S.: Detecting rumors through modeling information propagation networks in a social media environment. IEEE Transactions on Computational Social Systems **3**(2), 46–62 (2016)
36. Loftus, G.R.: Evaluating forgetting curves. Journal of Experimental Psychology: Learning, Memory, and Cognition **11**(2), 397 (1985)
37. Lopez, J., Scott, J.: Social structure Open. University Press (2000)
38. Luo, X., Xu, Z., Yu, J., Chen, X.: Building association link network for semantic link on Web resources. IEEE Trans. Autom. Sci. Eng. **8**(3), 482–494 (2011)
39. Luo, X., Xuan, J., Lu, J., Zhang, G.: Measuring the semantic uncertainty of news events for evolution potential estimation. ACM Transactions on Information Systems **34**(4), 24:1–24:25 (2016)
40. Makkonen, J.: Investigations on event evolution in tdt. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 Student Research Workshop - Volume 3, NAACLstudent ’03, pp. 43–48. Association for Computational Linguistics, Stroudsburg (2003)
41. Margalit, M., Leyser, Y., Avrahm, Y., Lewy-Osin, M.: Social-environmental characteristics (family climate) and sense of coherence in kibbutz families with disabled and non-disabled children. European Journal of Special Needs Education **3**(2), 87–98 (1988)
42. McGuinness, D.L., Van Harmelen, F., et al.: Owl Web ontology language overview. W3C Recommendation **10**(2004-03), 10 (2004)
43. Mei, Q., Zhai, C.: 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), KDD ’05, pp. 198–207. ACM, New York (2005)
44. Michelle, L., Roehm, A.M.T.: When will a brand scandal spill over, and how should competitors respond? J. Mark. Res. **43**(3), 366–373 (2006)
45. Milgram, S.: The small world problem. Psychology Today **2**(1), 60–67 (1967)
46. Mussweiler, T., Strack, F.: Hypothesis-consistent testing and semantic priming in the anchoring paradigm: a selective accessibility model. J. Exp. Soc. Psychol. **35**(2), 136–164 (1999)
47. Myers, J.L., O’Brien, E.J.: Accessing the discourse representation during reading. Discourse Processes **26**(2-3), 131–157 (1998)
48. Olick, J.K., Robbins, J.: Social memory studies: From collective memory to the historical sociology of mnemonic practices. Annu. Rev. Sociol. **24**(1), 105–140 (1998)
49. Papadimitriou, P., Dasdan, A., Garcia-Molina, H.: Web graph similarity for anomaly detection. Journal of Internet Services and Applications **1**(1), 19–30 (2010)
50. Radinsky, K., Bennett, P.N.: Predicting content change on the web. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), WSDM ’13, pp. 415–424. ACM, New York (2013)
51. Radinsky, K., Davidovich, S., Markovitch, S.: Learning causality for news events prediction. In: Proceedings of the 21st International Conference on World Wide Web (WWW), WWW ’12, pp. 909–918. ACM, New York (2012)
52. Radinsky, K., Horvitz, E.: Mining the Web to predict future events. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), WSDM ’13, pp. 255–264. ACM, New York (2013)
53. Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A.: Correlating financial time series with microblogging activity (2012)
54. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
55. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
56. Schriver, J.M.: Human behavior and the social environment. Allyn and Bacon (2004)
57. Slanzi, G., Balazs, J., Velásquez, J.D.: Predicting Web user click intention using pupil dilation and electroencephalogram analysis. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 417–420. IEEE (2016)
58. de Sola Pool, I., Kochen, M.: Contacts and influence. Social Networks **1**(1), 5–51 (1978)
59. Stavrianou, A., Andritsos, P., Nicoloyannis, N.: Overview and semantic issues of text mining. ACM SIGMOD Record **36**(3), 23–34 (2007)
60. Tao, X., Zhou, X., Zhang, J., Yong, J.: Sentiment Analysis for Depression Detection on Social Networks. In: Advanced Data Mining and Applications - 12Th International Conference, ADMA 2016, Gold Coast, QLD, Australia, December 12-15, 2016, Proceedings, pp. 807–810 (2016)
61. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. arXiv:1206.3298 (2012)

62. Wang, C., Lu, J., Zhang, G.: Integration of Ontology Data through Learning Instance Matching. In: WI 2006. IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp. 536–539 (2006)
63. Wang, C., Lu, J., Zhang, G.: Mining key information of Web pages: a method and its application. Expert Syst. Appl. **33**(2), 425–433 (2007)
64. Wang, C., Xiao, Z., Liu, Y., Xu, Y., Zhou, A., Zhang, K.: Sentiview: sentiment analysis and visualization for internet popular topics. IEEE Transactions on Human-Machine Systems **43**(6), 620–630 (2013)
65. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature **393**(6684), 440–442 (1998)
66. Whitworth, B., Whitworth, A.P.: The social environment model: Small heroes and the evolution of human society. First Monday 15(11) (2010)
67. Xuan, J., Luo, X., Zhang, G., Lu, J., Xu, Z.: Uncertainty analysis for the keyword system of Web events. IEEE Transactions on Systems, Man, and Cybernetics: Systems **46**(6), 829–842 (2016)
68. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM ’11, pp. 177–186. ACM, New York (2011)
69. Yang, Y., Carbonell, J.G., Brown, R.D., Pierce, T., Archibald, B.T., Liu, X.: Learning approaches for detecting and tracking news events. IEEE Intell. Syst. **14**(4), 32–43 (1999)
70. Yuan, H., Yuan, K., Zhao, Z.: On Predicting Event Propagation on Weibo. In: 2017 International Conference on Service Systems and Service Management, pp. 1–6 (2017)
71. Zhao, H., Zhou, H., Yuan, C., Huang, Y., Chen, J.: Social discovery: Exploring the correlation among three-dimensional social relationships. IEEE Transactions on Computational Social Systems **2**(3), 77–87 (2015)
72. Zhong, N., Liu, J., Shi, Y., Yao, Y.: An interview with professor raj reddy on Web intelligence (WI) and computational social science (CSS). Web Intelligence **16**(3), 143–146 (2018)

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Junyu Xuan[1,2] · Xiangfeng Luo[1] · Jie Lu[2] · Guangquan Zhang[2]

Junyu Xuan
Junyu.Xuan@uts.edu.au

Jie Lu
Jie.Lu@uts.edu.au

Guangquan Zhang
Guangquan.Zhang@uts.edu.au

[1] School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai, China

[2] The Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Sydney, Australia